# 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

**Sommario**
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.

## Indice



## Introduzione

Nel mondo del calcio, la figura del data analyst e data specialist è fondamentale, in ogni squadra, per effettuare analisi tecnico-tattiche a 360 gradi: si lavora coi dati nel contesto della match analysis, che viene effettuata sia per preparare partite future che per analizzare partite passate, dello scouting e anche del calciomercato, contesto che tratterò anche in questo progetto.

### Obiettivo

Lo scopo di questo progetto consiste nel trovare il giocatore che più possa assomigliare per caratteristiche fisiche, tecniche e tattiche al difensore centrale dell'Inter Francesco Acerbi, ragionando in un'ottica per cui questo giocatore possa essere il suo sostituto in futuro nella rosa dell'Inter, e quindi un obiettivo di mercato della società.
Il progetto si limita a considerare i giocatori militanti nei cinque campionati di livello più alto secondo il ranking UEFA [1] aggiornato al 17/03/2024, ovvero Premier League inglese, Liga spagnola, Serie A italiana, Bundesliga tedesca e Ligue 1 francese.

### Struttura del report

Il report, e anche i notebook su cui viene svolta l'analisi, sono organizzati come segue:

- **Data Retrieval**: ricerca delle fonti e conseguente ottenimento dei dataset di partenza, o con un semplice download e caricamento su notebook python oppure con tecniche di scraping;
- **Data Preparation**: insieme di operazioni di concatenazione dei dataset, pulizia, restringimento, feature selection, merging;
- **Data Modelling**: dopo una fase di normalizzazione e pesatura, in questa sezione presento i risultati, ottenuti con tecniche specifiche e bilanciando opportunamente i dati;
- **Data Visualization**: 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.

## 1. Data Retrieval

Per ottenere i dati su cui lavorare, ho preso in considerazione innanzitutto un dataset trovato su Kaggle [2], che contiene le statistiche sui giocatori che militano nei principali 5 campionati europei nella stagione 2022/23, ed è composto da 2689 tuple.
Le statistiche elencate nel dataset sono molte. Infatti, oltre alle colonne 'anagrafiche'

- Player: nome del giocatore
- Nation: nazionalità
- Pos: posizione in campo (inteso come ruolo, quindi ad esempio difensore, ala sinistra, ala destra...)
- Squad: squadra nella stagione 22/23
- Comp: campionato in cui milita quella squadra
- Age: età
- Born: anno di nascita

sono presenti 116 attributi, ciascuno dei quali valuta una statistica riguardante la stagione del giocatore. Oltre alle statistiche sul numero di partite giocate, i minuti giocati, il numero di partite in cui il giocatore è stato schierato titolare e il numero di goal nella stagione, tutte le altre sono pesate per 90 minuti di gioco.
Ho poi integrato questo dataset con un altro, proveniente sempre dalla stessa fonte [3], che riguarda i calciatori sempre militanti in squadre dei top 5 campionati europei, ma questa volta nel corso della stagione 2021/2022.

Dopo aver scaricato i due dataset, ho notato che essi sono effettivamente pieni di dati e statistiche interessanti, ma che ne mancassero alcune fondamentali per l'analisi che voglio effettuare, in particolare altezza e peso del giocatore - in un ruolo come quello del difensore centrale, è difficile pensare di poter sostituire Acerbi, alto un metro e novantadue, con un giocatore sotto il metro e ottanta - ma anche piede preferito e una stima del valore economico del giocatore, necessaria per effettuare una stima realistica su un'eventuale operazione di mercato.
Tutte queste informazioni le ho trovate disponibili su SoFIFA [4], un sito che si occupa della valutazione dei giocatori nel più celebre gioco a tema calcistico, ovvero fc24 di EA Sports (quello che un tempo si chiamava semplicemente FIFA). Dopo aver selezionato i parametri che effettivamente mi servivano nel menu apposito, ho quindi proceduto con uno scraping tradizionale (i dati sono presentati in un semplice formato a tabella) utilizzando la libreria Selenium di Python. In questo modo ho ottenuto le informazioni sopra anticipate.
Inoltre, filtrando sempre nel menu apposito, ho potuto scegliere di selezionare ovviamente solo i giocatori dei top 5 campionati (tra l'altro in questo caso questo viene valutato nella stagione 2023/2024, quindi chiaramente a causa di movimenti di mercato o ritiri o retrocessioni i giocatori non per forza corrispondono esattamente a quelli presenti negli altri due dataset), ma soprattutto ho potuto selezionare come posizione in campo quella di DC, ovvero difensore centrale. Questo mi permettere di escludere tutti i terzini, destri e sinistri, che per ovvi motivi logici non sono candidati realistici alla sostituzione di Acerbi; operazione che sugli altri dataset, come vedremo in fase di data preparation, non era possibile.

## 2. Data Preparation

Nella fase di data preparation, che è stata di gran lunga la più dispendiosa in termini di tempo ed impegno, si sono susseguite varie operazioni.

### 2.1 Filtro il primo dataset e ne elimino i duplicati

Prima di tutto, una volta caricato il primo dataset (quello proveniente da Kaggle, e riguardante la stagione 22/23), effettuando una breve analisi sull'attributo 'Comp' (che contiene il campionato in cui milita il giocatore in quella stagione) noto che questo contiene effettivamente solo i 5 campionati che avevo scelto come riferimento, come mi aspettavo.
Valutando poi sempre nello stesso modo la colonna 'Pos', contenente la posizione e il ruolo del giocatore in campo, noto che ce ne sono diverse, tutte rappresentate da delle sigle di due o quattro lettere. Estraendo la riga corrispondente a Francesco Acerbi, il mio giocatore di riferimento, trovo che lui è classificato come DF, che probabilmente sta per la parola 'Defender'. Effettivamente, valutando i nomi dei giocatori presenti nel dataset e indicati con altri valori di 'Pos', vedo che questi sicuramente non sono difensori, e quindi non sono sostituti papabili di Acerbi, per cui filtro il dataset conservando solo i giocatori classificati come DF, restringendo di circa il 70% le righe del dataset, che ora risultano essere 825.
DF però non distingue tra difensori centrali e terzini destri o sinistri, ma racchiude in sé tutti i giocatori classificabili come appunto 'difensori', in una classificazione simile a quella utilizzata ad esempio nel fantacalcio classico. Chiaramente non ha senso per la nostra analisi sostituire Acerbi con, non so, un Theo Hernandez, quindi i terzini andrebbero eliminati. Per fortuna, come anticipato nella sezione precedente, ci verrà in soccorso il terzo dataset.

Successivamente, mi occupo delle righe duplicate; ovvero, nel mio contesto, le situazioni in cui un giocatore compare due volte all'interno del dataset: questo è possibile perché stiamo valutando la stagione 2022/23, all'interno della quale un giocatore può aver cambiato squadra nel mercato invernale o in altri periodi. Dato che il dataset riporta per ogni giocatore la squadra e il campionato, è chiaro che se questi cambiano si crea una nuova riga per quel giocatore, con le nuove squadre o campionati.
L'austriaco Stefan Posch, ad esempio, dopo aver giocato la sua prima partita nella stagione con l'Hoffenheim in Bundesliga,

si è trasferito al Bologna, in Serie A, a settembre 2022, per poi giocarci 14 partite.
Vado quindi, per le righe duplicate, semplicemente a sommare ogni statistica: ha senso, sia per quelle numeriche assolute (partite giocate, minuti...) sia per quelle relative a 90 minuti di gioco (entrambi i dati che sommo sono normalizzati allo stesso modo, quindi non è necessario che li vada a moltiplicare per poi sommarli e dividerli ancora). Per quanto riguarda la squadra e il campionato, decido di conservare quelli in cui il giocatore ha giocato più minuti. Questo è consistente, perché comunque nulla mi dice che se avessi conservato l'ultima squadra in cui il calciatore ha giocato, questa sarebbe stata la sua squadra attuale: la stagione corrente è la 23/24, e sarà valutata col terzo dataset.

### 2.2 Opero una feature selection

Per gestire il numero elevatissimo di colonne nel dataset originale (116, oltre alle 7 anagrafiche) opero una feature selection, in cui vado a decidere di non considerare ulteriormente alcune di queste colonne, e quindi il corrispondente attributo, in particolare basandomi sul report riassuntivo del dataset, disponibile utilizzando la libreria sweetviz.
In base ai risultati del report [5], che tra le altre cose mostra il numero di valori distinti e nulli per ogni attributo e la correlazione di Pearson tra quell'attributo e gli altri del dataset, decido di procedere eliminando:

- la colonna 'Pos', che ormai è costituita solo dalla stringa DF, dato che ho eliminato precedentemente le altre possibilità; e 'Rk', che contiene semplicemente l'ordine originale in cui sono catalogati i giocatori e che non è quindi di alcun interesse
- la colonna 'Born', che riporta la data di nascita. Avendo a disposizione anche la colonna 'Age' e presentandosi una chiara forte correlazione tra le due, come mostrato dalla matrice di correlazione nel report, opto per tenere la seconda
- le colonne che contengono dati estemporanei, che non spiegano gli attributi di un giocatore e su cui non è corretto giudicarli in un campione di partite così ristretto. Elimino quindi 'OG' (own goals - numero di autogoal), 'PKwon' (calci di rigore guadagnati), 'PKcon' (calci di rigore concessi)
- '90s' (numero di partite in cui il giocatore è stato in campo 90 minuti), perché è un dato ridondante per coefficiente di Pearson con altre features che invece conservo ('Min', 'Starts', 'MP')
- alcuni dati sulla distanza dei passaggi, che si presentano con alta correlazione tra loro e che non sono troppo rilevanti per un difensore centrale, come per esempio il numero di passaggi corti, medi o lunghi completati: mi accontento di tenere il numero generale di passaggi completati
- altre informazioni un po' troppo ridondanti su azioni difensive, sui tocchi palla, sulla conduzione del pallone, sui dribbling, o comunque non rilevanti per difensori centrali
- i dati in percentuale che hanno un'equivalente non in percentuale in un altro attributo. Per lo stesso principio procedo eliminando anche le medie.

Al termine di questa feature selection, il dataset si riduce ad un totale di 60 attributi. Ne sono stati rimossi quindi più della metà.

### 2.3 Integro col dataset della stagione precedente

Dopo aver effettuato la selezione sulle colonne del dataset della stagione 2022/2023, procedo ad integrarlo con quello della stagione 2021/2022, andando a considerare una sola riga complessiva per i giocatori che sono presenti in entrambi i dataset.
Dato che il dataset relativo ai campionati 21/22 non ha esattamente le stesse identiche colonne dell'altro, e dato che non avrebbe senso effettuare anche su di esso una procedura di feature seletion, semplicemente escludo dal secondo dataset tutte le colonne non presenti anche nel primo.
Per quanto riguarda le altre, se sono numeriche procedo semplicemente sommandole - con l'unica eccezione della colonna che riguarda l'età - sfruttando lo stesso ragionamento che ho utilizzato quando ho trattato i duplicati. Se sono categoriche, invece, conservo i valori più recenti, ovvero quelli che si trovano nel primo dataset.

Un'ultima considerazione riguarda il fatto che i giocatori militanti nelle squadre dei top 5 campionati europei in una stagione non sono esattamente gli stessi che ci militavano nella stagione precedente. Chiaramente, se un giocatore sarà presente nel dataset più recente ma non in quello più datato, sarà comunque considerato nel dataset integrato. Non vale il caso opposto, ma i giocatori esclusi saranno solamente quelli ritiratisi, senza squadra nella stagione 22/23, che si sono trasferiti in campionati 'minori' o che vi si trovano perché la loro squadra vi è stata retrocessa, e che quindi avevo deciso di non considerare fin dall'inizio. Non sono quindi casi che devo preoccuparmi di trattare.

### 2.4 Pulizia della colonna 'Player'

Procedo ora col pulire la colonna 'Player', contenente il nome del giocatore, la quale presenta un po' di problemi. Prima di tutto, cerco di standardizzarla in più possibile, considerando che poi sarà su quella che dovrò effettuare il join col dataset proveniente dallo scraping: perciò, comincio rimuovendo gli accenti.
Aspetto però a rimuovere la punteggiatura: ce ne sarà da togliere, come ad esempio i trattini tra i nomi o i cognomi, abbastanza comuni a seconda del paese di provenienza del giocatore. Devo però prima valutare i punti di domanda presenti nei nomi, che noto essere molto frequenti. Mi accorgo

che sono sostanzialmente in sostituzione di caratteri accentati particolari, come le c accentate, molto comuni nei cognomi slavi. Procedo quindi sostituendoli a mano con le relative consonanti non accentate.
Dopo aver completato questa operazione, controllo quali sono i segni di punteggiatura rimasti: sono effettivamente solo i trattini, come anticipato. Procedo rimpiazzandoli con uno spazio vuoto, cosa che farò anche per il dataset proveniente da SoFIFA, che comincio a trattare ora.

### 2.5 Caricamento del dataset ottenuto da SoFIFA

Comincio quindi ora a valutare il dataset ottenuto scrapando da SoFIFA. In questo caso non ho chiaramente bisogno di fare una feature selection, considerando che questa è stata fatta già in fase di scraping, avendo io potuto selezionare le colonne di cui avevo bisogno.
Devo però pulire il dataset: come anticipato nella sezione sul data retrieval, da questa fonte ho potuto selezionare tutti i difensori centrali, i 'CB' (central back), escludendo così i terzini. Il lato negativo della cosa è che però, nella colonna 'Player', oltre al nome del giocatore ho sempre anche la sigla CB, unita a volte ad altre sigle, che indicano le altre posizioni che il calciatore può ricoprire in campo, come ad esempio 'CDM' (central defensive midfielder). Elimino quindi tutte queste sigle dalla colonna.
Un'operazione di pulizia va effettuata anche su praticamente tutte le altre colonne:

- in Squad, dopo il nome della squadra compare sempre una stringa del tipo '\n 20xx $\sim$ 20xx'. dove i due 20xx indicano l'anno di inizio e fine del contratto con quella squadra. Elimino sempre questa stringa.
- in Height, ho l'altezza espressa prima in cm, e poi, separata da un backslash (\), in piedi. Elimino sia la scritta 'cm' che tutti i caratteri che vengono dopo, lasciando solo la misura in centimetri, che poi converto in numerica.
- in Weight ho praticamente la stessa situazione, col peso espresso prima in chili e poi, dopo un backslash, in libbre. Come prima, lascio solo il peso in kg che poi andrò a considerare come numerico.
- le entrate di Value sono invece impostate con questo formato: il simbolo dell'euro, seguito dal valore numerico del prezzo, e infine dalla scala, che può essere M per milioni o K per migliaia. Rimuovo quindi il primo e l'ultimo carattere, ed esprimo tutti i numeri in milioni, per poi convertire l'attributo ancora una volta in numerico.

Alla fine di queste operazioni, dal dataset in questione posso estrarre informazioni interessanti, come ad esempio la lista dei centrali difensivi con valore di mercato maggiore. Riporto i primi 5, tra cui abbiamo anche un altro interista, Alessandro Bastoni:

| Giocatore | Squadra | Età | Mln |
|---|---|---|---|
| Ruben Dias | Manchester City | 26 | 106.5 |
| Ronald Araujo | Barcellona | 24 | 93 |
| Eder Militao | Real Madrid | 25 | 80 |
| Alessandro Bastoni | Inter | 24 | 73.5 |
| Marquinos | PSG | 29 | 73.5 |

### 2.6 Effettuo il join tra i due dataset

Ora mi ritrovo quindi ad avere due dataset preparati e puliti: il primo è l'unione dei due scaricati da Kaggle; il secondo, invece, è quello ottenuto da SoFIFA.
Devo joinarli, operazione tutt'altro che banale, considerando che non ho un attributo identico e univoco tra un dataset e l'altro: purtroppo SoFIFA non mi offre la possibilità di scrapare la data di nascita, che sarebbe stata molto d'aiuto, e quindi l'unica soluzione è effettuare il join sulla colonna 'Player', presente in entrambi i dataset, che contiene il nome del giocatore. Nome che però è rappresentato in maniera diversa: mentre nel primo dataset è effettivamente formato da una stringa 'Nome Cognome', nel secondo il formato è 'N. Cognome': nel nostro caso, 'F. Acerbi'.
Inoltre, questa è una considerazione che non vale per tutti i giocatori. Ci sono calciatori, soprattutto brasiliani, portoghesi o spagnoli, il cui nome può essere scritto in maniera diversa a seconda della convenzione che si attua, che purtroppo non è la stessa nelle due fonti: il difensore centrale brasiliano della Juventus Gleison Bremer nel secondo dataset è indicato semplicemente come 'Bremer', omettendo quello che è il nome proprio, e come lui tanti altri.

Forte di queste considerazioni, la strategia adottata è stata la seguente: prima di tutto, ho joinato i due dataset sull'attributo 'Player', ma non solo quando le due colonne sono esattamente uguali, bensì quando ad essere uguali sono almeno le ultime 5 lettere della stringa, se la seconda parola della stringa, ovvero il cognome, è cositutita per l'appunto da almeno 5 caratteri.
In questo modo ovvio al problema riguardante i nomi propri, sia se indicati con l'abbreviazione o se proprio non indicati. Ciò mi porta però inevitabilmente a molte associazioni errate, per esempio tra Alex Ferrari e G. Ferrari (Gianmarco Ferrai), entrambi difensori italiani ma chiaramente due persone diverse, oppure tra Luca Caldirola e Lorenzo Pirola, che hanno addirittura cognomi diversi, ma con le ultime 5 lettere identiche. Queste situazioni sono state inevitabilmente valutate 'a mano' e poi corrette.

Con questa strategia, efficace nel caso di cognomi lunghi, si andavano però chiaramente a non considerare tutti i giocatori con cognomi più corti, che non trovavano l'elemento con cui effettuare il join. Avrei escluso giocatori del calibro di Virgil Van Dijk e Ruben Dias (4 lettere), o Kim Min Jae e Nathan Aké (3). Avrei anche escluso il giocatore che poi si scoprirà essere il più simile ad Acerbi. Perciò, tutta l'operazione di cui

sopra è stata ripetuta anche per cognomi (ultime parole nella stringa) lunghi 4 caratteri, poi 3, infine 2.
I dataset ottenuti in questo modo sono stati poi concatenati con quello che considerava le ultime 5 lettere, per avere un totale di 880 giocatori.

Mi ritrovo però ora pieno di colonne NaN, in particolare provenienti dal secondo dataset. Questo è consistente, perché il primo dataset contiene anche i terzini, che nel secondo dataset invece sono giustamente assenti, o anche dei difensori centrali che nella stagione 2022/2023 militavano in squadre appartenenti ai top 5 campionati, mentre in quella attuale no (come Rodrigo Becao, passato dall'Udinese, in Serie A, al Fenerbache, in Turchia).
Vado allora, dopo aver effettuato un controllo nel caso non avessi considerato altri passaggi invece necessari, ad eliminare le righe in cui sono presenti dei valori nulli. Mi ritrovo allora col mio dataset quasi finale, con 380 elementi, che corrispondono ai difensori centrali che militano nei top 5 campionati europei in questa stagione

## 2.7 Restringo il dataset finale, valutando Acerbi

Infine, dopo aver ottenuto il dataset praticamente definitivo, concludo con un'ultima valutazione 'realistica' e in realtà sostanzialmente anagrafica su che tipo di giocatore possa volere in sostituzione di Francesco Acerbi.
Al di là di caratteristiche tecniche, che valuteremo in fase di modelling, infatti, ci sono fattori anagrafici impossibili da ignorare, e che conviene affrontare prima di considerare i modelli, in modo da restringere preliminarmente il dataset.
Valutando l'età attuale di Acerbi, che ottengo dalla colonna 'Age', creata per ogni giocatore semplicemente aumentando di 1 l'età indicata nel dataset riguardante la stagione 22/23, ho infatti che Acerbi ha attualmente 36 anni, un'età molto avanzata per un giocatore di calcio, come evidenziato dalla distribuzione di età presente nel dataset

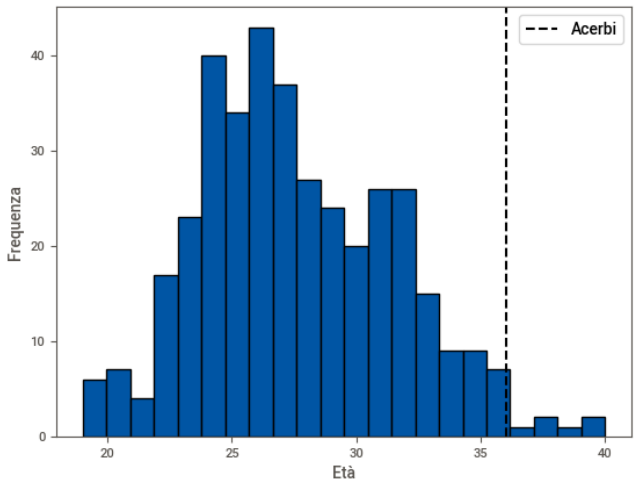



**Figura 1.** *Età dei giocatori nel dataset, in verticale in nero Francesco Acerbi*

Perciò, l'idea è quella di restringere il dataset considerando solo i giocatori di età almeno inferiore a quella di Acerbi: non è infatti realistico che l'Inter cerchi sul mercato un suo sostituto che sia 'più vecchio' anagraficamente, essendo lui già avanti con gli anni nel mondo del calcio professionistico. Ad ogni modo, come si nota dal grafico, chiaramente con questo restringimento non si vanno ad escludere molti giocatori, ma solo 16.
Inoltre, decido anche di escludere i giocatori che abbiano giocato meno di 20 partite di campionato sommando gli ultimi due anni, su un totale di partite disponibili di 76 nel caso di Premier League, Liga e Serie A, e 68 nel caso di Bundesliga e Ligue1. Questo decido di farlo sia perché non avrei molti dati a disposizione su questi giocatori per rapportarli alle caratteristiche di Acerbi, e questo influerenzerebbe negativamente la fase di modelling, sia perché, sempre realisticamente e in un'ottica di mercato, l'Inter come sostituto di un pilastro della difesa come Acerbi andrà a cercare un giocatore già relativamente pronto ed esperto, non incline ad infortuni e subito schierabile titolare.

## 2.8 Ultime operazioni

Infine, come ultime operazioni prima di poter considerare il dataset come definitivo, e poterlo quindi utilizzare nella fase di modelling, sono stati effettuati i seguenti passaggi:

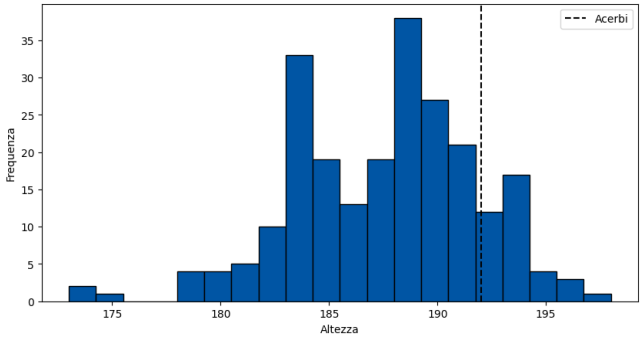
- trasformazione in interi delle colonne che contengono attributi che non sia necessario avere come float, ovvero 'MP' (Matchs Played), 'Starts' (partite iniziate da titolare) 'Min' (minuti giocati) e 'Age';
- pesatura dell'attributo 'Goals' all'interno dei 90 minuti, in modo che si presenti come tutti gli altri attributi riguardanti le statistiche sulla stagione del giocatore e che non riguardano solamente il tempo trascorso in campo;
- ordinamento del dataset in ordine alfabetico per il cognome, e riassegnamento degli indici corrispondenti ai giocatori in base a questo ordinamento;
- riordinamento delle colonne;
- gestione dei duplicati, creatosi in particolare unendo i vari dataset creati in fase di joining in base alla lunghezza del cognome dei giocatori. Giocatori coi cognomi inferiori alle cinque lettere possono apparire come duplicati. Questi vengono semplicemente eliminati, mentre altri sono gestiti a mano.

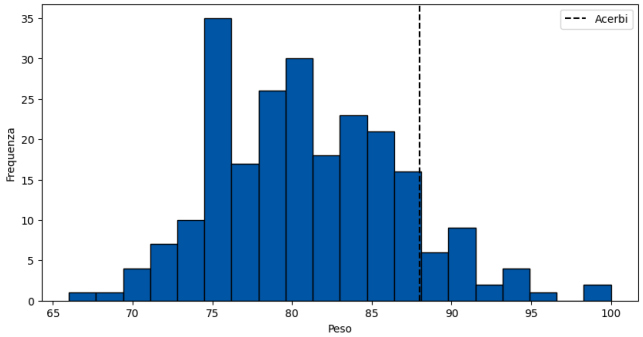
Per riassumere, il dataset finale sarà quindi costituito da 233 difensori centrali, Acerbi compreso, militanti nei top 5 campionati europei - nei quali hanno disputato almeno 20 partite negli ultimi due anni - e più giovani di Acerbi. Non ho alcun dato mancante.

## 3. Data Modelling

Per approcciare la fase di data modelling, prima di tutto ho bisogno di catalogare gli attributi di ogni giocatore in quattro categorie: quelli prettamente anagrafici (che sono sostanzialmente le stringhe, che risultano inutili se non a scopo informativo, non potendo essere quantificate numericamente), quelli che riguardano caratteristiche offensive, quelli che invece si focalizzano sull'aspetto difensivo del gioco (o comunque quello in assenza di possesso del pallone) e infine quelli, classificati come 'altro', che sono numerici ma che riguardano più il giocatore in sé che le sue prestazioni: abbiamo quindi sostanzialmente tutti i dati provenienti dallo scraping, ovvero altezza, peso, valore economico, piede preferito, e anche dati come età o partite giocate.
Alla fine di questa procedura di catagolazione ho 32 statistiche prettamente offensive, 20 difensive, e 8 'personali', da trattare diversamente dalle altre.

Prima di procedere con una procedura efficace che tenga in considerazione le statistiche prestazionali del giocatore, quindi, inizio come fatto in precedenza con un'analisi su Acerbi, il giocatore target, effettuata rispetto agli altri giocatori presenti nel dataset.
Estraggo quindi la sua altezza (192cm), il suo peso (88kg) e il suo piede preferito (sinistro). Noto che sia per quanto riguarda altezza che peso, Acerbi si trova ad essere fisicamente molto prestante, rispetto alla distribuzione del dataset. E anche per quanto riguarda il piede preferito, essendo l'ex Lazio mancino, si trova in una categoria di calciatori che risulta meno popolosa rispetto a quella dei destri naturali (nel mio dataset, il rapporto è di 66 a 167).



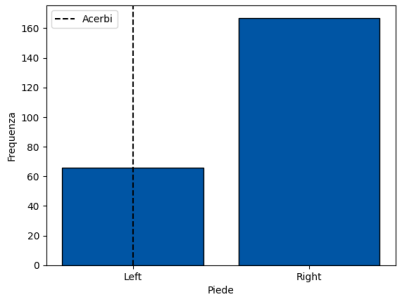




**Figura 2.** *Altezza, peso e piede preferito dei giocatori nel dataset, in verticale in nero tratteggiato dove si posiziona Francesco Acerbi*

Queste peculiarità di Acerbi non possono essere ignorate: la prestanza fisica è chiaramente una caratteristica importante in un difensore centrale, determinante ad esempio nei colpi di testa e nei contrasti, e anche il piede preferito è fondamentale in fase di impostazione, soprattutto in un gioco fluido e totale come quello dell'Inter di Inzaghi. D'altro canto, non posso e non voglio nemmeno escludere tutti i giocatori destri, oppure inferiori ad una certa altezza o ad un certo peso, perché magari queste loro differenze con Acerbi potrebbero essere compensate da similarità evidenti che si possono riscontrare tra i dati più tecnici.
Per questo, decido di assegnare ad ogni giocatore uno score di similarità con Acerbi, che valuti l'altezza, il peso e il piede preferito del singolo calciatore. Questo score verrà poi integrato con quello che otterrò a partire dai dati prestazionali difensivi e offensivi.

Sia per quanto riguarda l'altezza che per quanto riguarda il peso, definisco uno score in base ad una scala esponenziale che valuta la differenza di altezza e peso di ogni giocatore con Acerbi. Questa scala prevede un parametro alpha variabile tra 0 e 1, che nel nostro caso è fissato a 0.25 ma può essere modificato in base a quanta importanza e a quanto peso nella valutazione finale si vuole dare a queste caratteristiche fisiche rispetto alle altre, più tecnico-tattiche.
Lo score per il piede preferito, invece, è definito semplicemente pari a 2/3 se il piede preferito è lo stesso di Acerbi, ovvero il sinistro, e 1/3 se è invece il destro. Anche questi valori sono ovviamente variabili.
Lo score finale che racchiude queste tre caratteristiche è semplicemente dato dal prodotto di questi tre, che viene però riscalato con la tecnica min-max, in modo che giaccia nell'intervallo $[0, 1]$.

A questo punto, mi concentro sui dati 'prestazionali' dei giocatori, ovvero quelli che ho raggruppato e diviso in difensivi ed offensivi. Dopo aver normalizzato i dati, che chiaramente giacciono in intervalli diversi ed hanno inevitabilmente distribuzioni diverse, operazione che viene eseguita utilizzando ancora una volta la strategia min-max, devo pesarli adeguatamente.

Infatti, come anticipato all'inizio di questa sezione, ho 32 statistiche che riguardano l'aspetto offensivo, o comunque la fase di possesso palla, del gioco - tiri, tocchi palla, passaggi tentati, avanzamenti palla al piede... - e 'solamente' 20 riguardanti invece l'aspetto difensivo, o perlomeno che si sviluppano col pallone in possesso dell'avversario - tackles, blocchi, intercettazioni, cartellini gialli, duelli aerei...
Devo cercare di bilanciare questa situazione: consapevole che è verissimo che nel calcio moderno la fase d'impostazione è fondamentale anche a partire dai difensori centrali (ed infatti in fase di feature selection ho preferito non rimuovere molte statistiche offensive, anche se avrei potuto), la fase difensiva continua comunque ad essere la priorità, e quindi non può pesare meno di quella offensiva nel calcolo finale. Una volta normalizzati i dati decido quindi, ed è una scelta abbastanza arbitraria, di aggiungere ai dati difensivi un coefficiente moltiplicativo di 3.2 ($\frac{32}{20} \cdot 2$), in modo che nel complesso le caratteristiche difensive di un giocatore pesino esattamente il doppio rispetto a quelle offensive. Ripeto che questa scelta è comunque totalmente soggettiva, avrei potuto utilizzare un coefficiente maggiore o minore in base ad un diverso bilanciamento che avrei potuto voler dare alle caratteristiche dei giocatori, oppure un singolo coefficiente per ogni statistica, o per raggruppamento di statistiche, se fosse stato fatto in maniera più granulare (ad esempio, dividendo in dati riguardanti passaggi, contrasti, cartellini etc).

Ora, avendo normalizzato e pesato i dati, posso effettivamente calcolare la similarità tra tutti i giocatori del mio dataset e il giocatore di riferimento, Acerbi. Per farlo, utilizzo la tecnica della cosine similarity [6], che tratta le tuple del dataset come vettori in uno spazio a più dimensioni, le quali corrispondono agli attributi numerici. Dopodiché, la cosine similarity prevede l'utilizzo della formula

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\cdot\|\vec{b}\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

per calcolare l'angolo che si trova tra il vettore 'Francesco Acerbi' e ogni altro vettore, uno per volta. In base all'ampiezza di quest'angolo viene stabilita la similarità: più è piccolo, più i vettori sono vicini e quindi i giocatori simili.

A questo punto, per riassumere, mi trovo in una situazione in cui ho calcolato uno score che possiamo definire 'fisico', o 'personale', che deriva dalle informazioni su altezza, peso e piede preferito (che quindi non dipendono dalle prestazioni e dal rendimento del calciatore nel corso delle due stagioni in esame); e uno definibile come 'prestazionale', calcolato con la tecnica della similarità del coseno, che si basa sui dati normalizzati su statistiche difensive e offensive raccolte nelle ultime due stagioni e pesate sui 90 minuti.
Voglio ora combinare questi due score in un unico score, che sia complessivo di tutti i miei dati numerici. Per farlo, decido di sommarli assegnandogli però un peso: lo score riguardante i fattori 'personali' varrà solamente 1/10 dello score complessivo, mentre i restanti 9/10 che costituiranno lo score finale sono quelli provenienti dall'addendo 'prestazionale'. Questa scelta così all'apparenza sbilanciata può sembrare che possa favorire i dati relativi alle ultime due stagioni rispetto a quelli fisici e propri del giocatore, ma ciò in realtà non risulta per come sono stati definiti gli score parziali. Anche qui, comunque, la scelta su come pesare gli scores può variare in base al maggior risalto che si può voler dare ai dati fisici rispetto a quelli prestazionali, o viceversa.
Ottengo comunque quindi un valore unico di similarità di ogni giocatore con Acerbi, valore che appartiene all'intervallo $[0,1]$. Più questo valore è vicino a uno, più il giocatore gli è simile.

Ora, semplicemente ordinando il dataset in maniera decrescente per i valori dello score complessivo definitivo, estraggo i cinque difensori centrali che, secondo i miei parametri, risultano essere quelli più 'vicini' ad Acerbi. Nella tabella sottostante riporto qualche informazione su questi potenziali acquisti:

| Giocatore | Squadra | Campionato |
|---|---|---|
| Stefan Bell | Mainz | Bundesliga |
| Maxence Lacroix | Wolfsburg | Bundesliga |
| Matthijs de Ligt | Bayern Monaco | Bundesliga |
| Robin Koch | Eintracht Frankfurt | Bundesliga |
| Nico Schlotterbeck | Borussia Dortmund | Bundesliga |

| Nazione | Età | Altezza | Peso | Piede | Score |
|---|---|---|---|---|---|
| GER | 32 | 192 | 88 | DX | 0.917 |
| FRA | 24 | 190 | 88 | DX | 0.910 |
| NED | 24 | 189 | 89 | DX | 0.895 |
| GER | 27 | 191 | 85 | DX | 0.894 |
| GER | 24 | 191 | 86 | SX | 0.893 |

**Tabella 1.** *I cinque giocatori che ottengo come più simili ad Acerbi*

Questi giocatori risultano essere il tedesco Stefan Bell, del Mainz, il francese Maxence Lacroix, del Wolfsburg, l'olandese Matthijs de Ligt, del Bayern, il tedesco Robin Koch, dell'Eintracht, e infine un altro tedesco, Nico Schlotterbeck, del Borussia Dortmund. Per una curiosa coincidenza, nonostante sono stati presi in esame i calciatori di ben 5 campionati, tutti i giocatori ottenuti da questa analisi militano in uno solo di questi, quello tedesco, la Bundesliga.

## 4. Data Visualization

Effettuo una prima analisi sulle caratteristiche che ho definito come 'fisiche' o 'personali. Mostro prima di tutto uno scatterplot che riguarda l'intero dataset dei centrali di difesa - quello, per intenderci, ottenuto dopo la fase di data preparation e su cui si è applicato il modello.

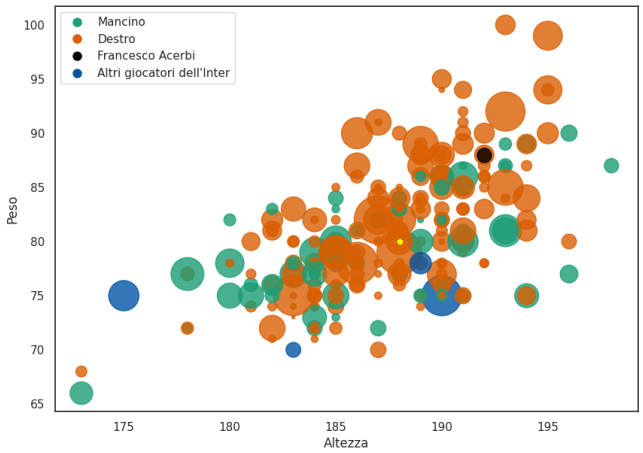



**Figura 3.** *Scatterplot altezza vs peso, con dimensione data dal valore di mercato*

In questo scatterplot, dove sono indicati sugli assi le altezze e i pesi dei giocatori, e dove la dimensione del pallino corrispondente al giocatore è proporzionale al suo valore di mercato, sono rappresentati tutti i difensori centrali di riferimento. Acerbi è evidenziato dal colore nero, mentre gli altri giocatori dell'Inter sono visualizzati in blu. In verde abbiamo i mancini come Acerbi, in arancione i destri. In giallo la posizione corrispondente all'altezza e il peso mediani.
Possiamo notare, come comunque già evidenziato in precedenza, come Acerbi si discosti abbastanza dalle caratteristiche mediane, attorno al quale sostanzialmente lo scatterplot si dirama con una distribuzione normale, con qualche outlier sia positivo che negativo.

Vediamo ora un secondo scatterplot, che stavolta prende in considerazione solo i giocatori selezionati per essere i 5 più adatti a sostituire Acerbi:

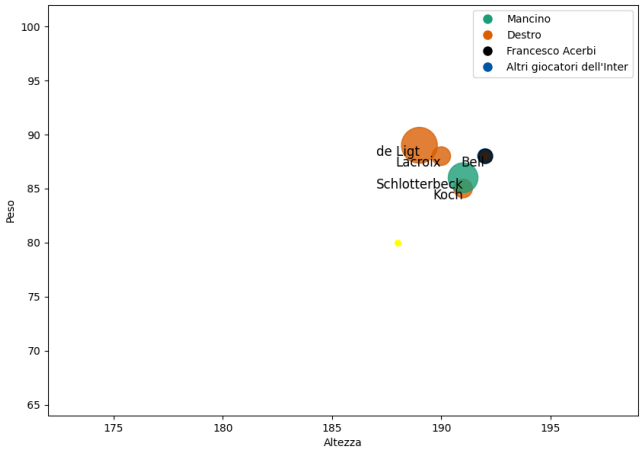



**Figura 4.** *Scatterplot altezza vs peso ristretto ai 5 giocatori scelti*

Da questo grafico prima di tutto notiamo come la distribuzione delle posizioni sia nettamente più accostata alla posizione di Acerbi di quanto lo sia quella nel caso dell'intero dataset, nonché come sia più distante dalla posizione mediana rappresentata in giallo. Inoltre, notiamo come nessun giocatore tra quelli scelti sia più alto di Acerbi, solo Bell lo eguaglia. Bell tra l'altro ha proprio le stesse caratteristiche fisiche di Acerbi: infatti ha il pallino esattamente sovrapposto al suo, e questo è consistente col fatto che sia considerato il giocatore più simile, anche se comunque ovviamente questo grafico non tiene conto delle statistiche 'prestazionali', quindi non per forza sarebbe stato così.
La rilevanza delle altre caratteristiche si nota per esempio dal fatto che De Ligt si posiziona palesemente più lontano da Acerbi rispetto a quanto non lo sia Lacroix, ma gli è comunque davanti come scelta per sostituirlo, evidentemente a causa degli altri dati.
Schlotterbeck, invece, è l'unico giocatore ad essere mancino come Acerbi, caratteristica che pone a suo favore.

Andiamo ora a presentare un'altra coppia di scatterplot, che come in precedenza anche questa volta mostrano il variare della situazione prima all'interno dell'intero dataset e successivamente limitata ai cinque giocatori di riferimento (sei con Acerbi).
In questo caso, lo scatterplot vuole evidenziare sull'asse delle ascisse il valore che è stato assegnato allo score cosiddetto fisico o personale; mentre su quello delle ordinate lo score di riferimento per le statistiche prestazionali. La legenda e il criterio con cui è stata scelta la dimensione dei pallini sono esattamente gli stessi che sono stati utilizzati in precedenza. Viene solamente aggiunto un pallino rosso ad indicare non le posizioni mediane, bensì quelle medie: in questo caso la distribuzione non può essere assimilata ad una normale a due dimensioni, come avveniva precedentemente con peso e altezza, per cui la posizione media si discosta da quella mediana. In particolare, si discosta molto di più per quanto riguarda lo score 'personale', il che è assolutamente pronosticabile considerando che è stato costruito facendo uso di funzioni esponenziali.

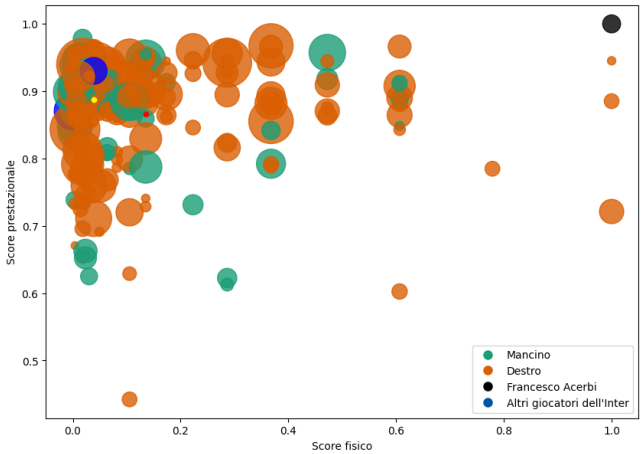



**Figura 5.** *Scatterplot Score fisico vs Score prestazionale*

Analizzando il grafico notiamo come, avendo scelto una scala

esponenziale per definire gli score di altezza e peso, lo score 'fisico' si discosta molto di più rispetto al valore ideale pari ad 1, raggiunto solo da Acerbi ed altri tre giocatori, rispetto a quanto si discosti invece quello prestazionale. Questo conferma quanto fosse necessario pesare poi gli score nel momento di definire quello definitivo.

Vado allora a riportare i dati riguardanti i giocatori che sarebbero stati scelti con criteri differenti: prima nel caso avessimo considerato solo i dati di altezza, peso e piede preferito; poi, se invece avessimo trascurato tutti quelli per soffermarci solo sulle statistiche offensive e difensive provenienti dal primo dataset:

| Giocatore | Squadra | Campionato | Score |
|---|---|---|---|
| Martin Erlic | Sassuolo | Serie A | 1 |
| Lewis Dunk | Brighton | Premier | 1 |
| Stefan Bell | Mainz | Bundesliga | 1 |
| Sepp van der Berg | Mainz | Bundesliga | 0.77 |
| Maxence Lacroix | Wolfsburg | Bundesliga | 0.6 |

**Tabella 2.** *I 5 giocatori più simili ad Acerbi considerando solo il primo score*

| Giocatore | Squadra | Campionato | Score |
|---|---|---|---|
| Lilian Brassier | Brest | Ligue 1 | 0.9779 |
| Matthijs de Ligt | Bayern | Bundesliga | 0.9665 |
| Maxence Lacroix | Wolfsburg | Bundesliga | 0.9663 |
| Robin Koch | Eintracht | Bundesliga | 0.9659 |
| Robin Knoche | Union | Bundesliga | 0.9656 |

**Tabella 3.** *I 5 giocatori più simili ad Acerbi considerando solo il secondo score*

Si nota come in queste tabelle vengano mostrati alcuni giocatori che effettivamente 'sopravviveranno' in quello che è poi il risultato finale, mentre altri i quali, essendo evidentemente simili ad Acerbi solo per quanto riguarda il primo score piuttosto che il secondo (o viceversa) non vengono poi riportati nella top 5 finale. Un esempio eclatante è costituito, riferendoci alla seconda tabella, dal francese Lilian Brassier, il quale risulta essere il giocatore che a livello di prestazioni si avvicina di più ad Acerbi, con anche un discreto vantaggio rispetto al secondo classificato, ma che con un'altezza e un peso rispettivamente di 186cm e 78kg risulta essere molto più minuto del centrale dell'Inter. Lilian Brassier, comunque, facendo valere la forte similarità a livello di statistiche offensive e difensive, si posiziona al sesto posto per quanto riguarda lo score generale, quindi appena fuori dalla top 5.
Invece, l'unico giocatore ad essere presente nelle top 5 di entrambi gli score provvisori e ad essere anche presente nella top 5 complessiva è un altro francese, Maxence Lacroix, che nella complessiva si posiziona al secondo posto.

Presento ora, come fatto in precedenza, un grafico analogo, dove però considero solo i giocatori selezionati come simili, e che ne evidenzia le posizioni rispetto ad Acerbi e alle posizioni medie e mediane:

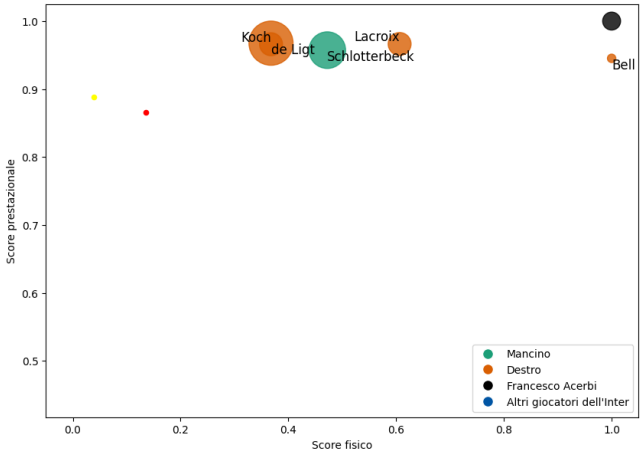



**Figura 6.** *Scatterplot Score fisico vs Score prestazionale ristretto ai 5 giocatori scelti*

## Conclusioni e sviluppi futuri

Analizzando i risultati che abbiamo ottenuto, si ha che risulta conveniente andare a cercare il sostituto di Acerbi nel campionato tedesco, campionato che effettivamente è molto fisico e molto adatto a difensori con caratteristiche simili alle sue.
Ragionando però, come si è sempre detto, in un'ottica realistica di mercato, un fattore che dobbiamo assolutamente tenere in considerazione è il valore economico che il giocatore assume appunto sul mercato. Mostro allora, per ogni giocatore tra i 10 più simili ad Acerbi, il valore di mercato. Approfitto tra l'altro di questa tabella per notare come la Bundesliga la faccia effettivamente da padrona anche estendendoci ai primi 10 giocatori (occupando 7 posizioni), mentre come invece in Italia non ci sia nessun giocatore particolarmente adatto a ricoprire il ruolo di Acerbi, quindi andrebbe in ogni caso scelto un giocatore che dovrà adattarsi al campionato.

| Giocatore | Squadra | Valore (€) |
|---|---|---|
| Stefan Bell | Mainz | 2.1 Mln |
| Maxence Lacroix | Wolfsburg | 15.5 Mln |
| Matthijs de Ligt | Bayern Monaco | 57.5 Mln |
| Robin Koch | Eintrach Francoforte | 15.5 Mln |
| Nico Schlotterbeck | Borussia Dortmund | 39.5 Mln |
| Lilian Brassier | Brest | 10.5 Mln |
| Kevin Vogt | Union Berlino | 5 Mln |
| Edmond Tapsoba | Bayer Leverkusen | 33 Mln |
| Kevin Danso | Lens | 27 Mln |
| Axel Disasi | Chelsea | 23 Mln |

Soffermandoci sull'aspetto economico, invece, anche lì andranno fatte valutazioni: Bell, che è il giocatore ideale, è anche molto economico, addirittura il più economico tra i primi 10. Si parla però di un giocatore che va per i 33 anni e che quindi, seppur sia più giovane di Acerbi, ha un'età comunque elevata

e che giustifica il prezzo. Andrebbero quindi fatte valutazioni societarie: è chiaro che per giocatori più giovani e militanti in top club, come de Ligt e Schlotterbeck, entrambi del '99, il prezzo si impenni, ma è pur sempre vero che, rivendendoli dopo qualche anno, il costo a bilancio potrebbe venire ammortizzato, cosa più difficile per calciatori over 30.
In questo senso effettivamente il già citato Lilian Brassier potrebbe essere il profilo più interessante. Ok, è solo sesto nella 'graduatoria', ma primo per i valori prestazionali, anche lui del '99 e relativamente economico.

In ogni caso, comunque, come ripetuto più e più volte all'interno del report, le valutazioni andrebbero fatte in base alle esigenze della società, della squadra e dell'allenatore, che spesso convergono ma a volte possono anche essere in contrasto tra loro.
Tutte le strategie effettuate e portate avanti nel progetto, e tutti i coefficienti numerici assegnati all'interno dei calcoli, non sono altro che frutto di scelte comunque arbitrarie, e modificabili in caso di direttive diverse. In futuro sarebbe probabilmente interessante sviluppare in un progetto del genere un'interfaccia che permetta all'utente - che potrebbe essere uno scout o comunque una figura dirigenziale all'interno dell'Inter - di andare a selezionare manualmente quali possano essere gli aspetti fisici, tecnici o tattici (o anche, perché no, economici) che più si vogliono tenere in considerazione nell'effettuare questa analisi.
Un'interfaccia del genere potrebbe andare per esempio ad escludere a priori certi giocatori, come magari i destri naturali, o quelli che costano più di 20 milioni, o comunque a dare più peso a statistiche o gruppi di statistiche che possono essere considerate più rilevanti di altre, come ad esempio quelle difensive, o più nello specifico tutte quelle sui tackles, o ancora più nello specifico quella sui tackles limitata solo alla trequarti offensiva. Quello che viene riportato all'interno di questo report è solo uno dei tanti possibili scenari che, con un'interfaccia il più possibile intuitiva, potrebbero venire volta per volta selezionati senza andare in alcun modo ad intaccare direttamente il codice.

## Riferimenti bibliografici


[1] Ranking UEFA:
*https://www.uefa.com/nationalassociations/uefarankings/country/#/yr/2024*

[2] Primo dataset ottenuto da Kaggle:
dataset 2022/2023

[3] Secondo dataset ottenuto da Kaggle:
dataset 2021/2022

[4] SoFIFA

[5] Il report si può consultare accedendo ai file del notebook

[6] Cosine similarity:
*https://www.machinelearningplus.com/nlp/cosine-similarity/*