I dati nel mondo del calcio

Lorenzo Lecce (n° matricola 830881) l.lecce@campus.unimib.it

Andrea Lucini Paioni (n° matricola 826578) a.lucinipaioni1@campus.unimib.it

Abstract

L'obiettivo di questo project work è stato di progettare un database relativo ad informazioni sui calciatori di alcuni dei campionati principali al mondo per la stagione 2022/23: il tutto è stato creato a partire da tre siti (**Capology**, **Transfermarkt** e **FBREF**) che raccolgono statistiche e informazioni relative al mondo del calcio.

Introduzione

Nel corso degli ultimi anni, il mondo del calcio ha vissuto importanti cambiamenti sotto diversi punti di vista, da quello economico a quello tecnico-tattico. Col passare del tempo, è cresciuta l'importanza che hanno i dati e l'analisi di partite, allenamenti, profili dei giocatori nel migliorare i risultati, con squadre che utilizzano sempre più sofisticati sistemi di raccolta e analisi dei dati per ottenere un vantaggio competitivo, per ridurre al minimo gli infortuni, rendere al massimo delle proprie potenzialità, ricercare profili mirati in base alle esigenze della squadra.

Le squadre sono sempre più attente alla preparazione fisica e all'alimentazione dei giocatori, vengono inserite nelle società sempre più figure altamente specializzate (come data analyst e scout), e utilizzano sofisticati sistemi di scouting per individuare in anticipo i talenti del futuro, così da poter avere un vantaggio economico sui competitor.

Inoltre, l'avvento del Covid-19 ha avuto un impatto sulla gestione tecnica e finanziaria delle società (essendo pur sempre aziende, esse hanno risentito non poco della crisi che ha portato il Covid in tutti i settori), costringendo le federazioni a dover adottare misure per garantire la sicurezza di spettatori e calciatori.

In particolare, alcune società più di altre sono contraddistinte dall'utilizzo massiccio di dati nel compiere scelte strategiche sul mercato, nella gestione dei giocatori e degli infortuni, nello studio degli avversari. Tra queste, l'AC Milan, società tra le più vincenti della storia dello sport, ha sconvolto il mondo del calcio con alcune drastiche decisioni strategiche compiute al termine della stagione 2022/23: licenziare Paolo Maldini, storico dirigente nonché ex-leggenda dello sport e per 20 anni calciatore proprio del Milan, per passare ad un approccio basato in modo preponderante sui dati, a partire dalle decisioni prese sul mercato.

Solo il tempo dirà se questa decisione, fortemente voluta dalla proprietà americana, sarà stata un'ottima o una pessima idea. Ma nel frattempo, abbiamo provato ad immaginare, in piccolo, come possa essere questo approccio al cosiddetto "*moneyball*", come è stato definito dal giornalismo sportivo.

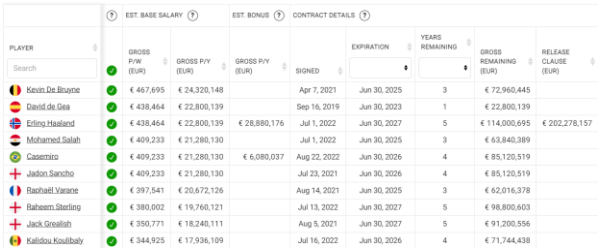
Lo scopo di questo studio, dunque, è di ricreare un dataset in cui vengono raccolte informazioni relative ad un gran numero di calciatori, come se fosse uno degli strumenti a disposizione dei dirigenti del Milan (pur con le dovute differenze date dal fatto che i dati forniti gratuitamente online sono sicuramente meno e coinvolgono un numero minore di calciatori rispetto a quelli che sono a disposizione di una società di calcio come il Milan).

Acquisizione dei dati

La prima fase consiste nell'acquisizione dei dati di interesse. Sono tre le fonti principali da cui abbiamo acquisito i dati relativi ai singoli giocatori:

Capology: un sito che si occupa della raccolta di dati relativi agli aspetti finanziari del mondo del calcio, sia a livello di campionato che per ogni singola società:

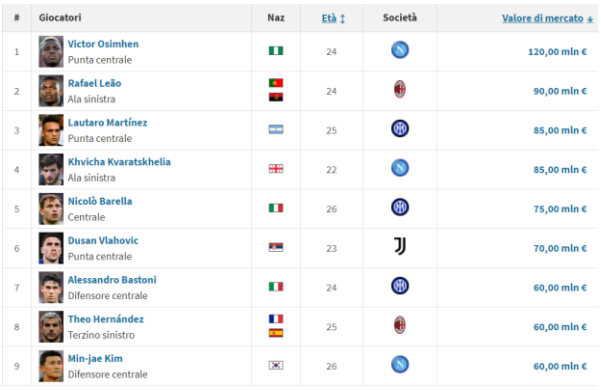
- Tutti i differenti aspetti relativi ai contratti dei calciatori (dagli stipendi, alla durata dei contratti, ai bonus inseriti nei contratti, fino alla presenza di clausole rescissorie;
- Gli aspetti relativi alle finanze delle società di calcio (come il bilancio, i ricavi, le spese…)
- I dati relativi ai trasferimenti (dunque per le spese e i ricavi dati dal calciomercato).



PLAYER		EST. BASE SALARY		EST. BONUS	CONTRACT DETAILS				
		GROSS P/W (EUR)	GROSS P/Y (EUR)	GROSS P/Y (EUR)	SIGNED	EXPIRATION	YEARS REMAINING	GROSS REMAINING (EUR)	RELEASE CLAUSE (EUR)
Kevin De Bruyne	✓	€ 467,695	€ 24,320,148		Apr 7, 2021	Jun 30, 2025	3	€ 72,960,445	
David de Gea	✓	€ 438,464	€ 22,800,139		Sep 16, 2019	Jun 30, 2023	1	€ 22,800,139	
Erling Haaland	✓	€ 438,464	€ 22,800,139	€ 28,880,176	Jul 1, 2022	Jun 30, 2027	5	€ 114,000,695	€ 202,278,157
Mohamed Salah	✓	€ 409,233	€ 21,280,130		Jul 1, 2022	Jun 30, 2025	3	€ 63,840,389	
Casemiro	✓	€ 409,233	€ 21,280,130	€ 6,080,037	Aug 22, 2022	Jun 30, 2026	4	€ 85,120,519	
Jadon Sancho	✓	€ 409,233	€ 21,280,130		Jul 23, 2021	Jun 30, 2026	4	€ 85,120,519	
Raphaël Varane	✓	€ 397,541	€ 20,672,126		Aug 14, 2021	Jun 30, 2025	3	€ 62,016,378	
Raheem Sterling	✓	€ 380,002	€ 19,760,121		Jul 13, 2022	Jun 30, 2027	5	€ 98,800,603	
Jack Grealish	✓	€ 350,771	€ 18,240,111		Aug 5, 2021	Jun 30, 2027	5	€ 91,200,556	
Kalidou Koulibaly	✓	€ 344,925	€ 17,936,109		Jul 16, 2022	Jun 30, 2026	4	€ 71,744,438	

Figura 1: esempio di dataset di Capology degli stipendi della Premier League, stagione 2022/23

Transfermarkt: un sito fondato nel 2000 che si occupa della raccolta di statistiche e dati relativi al mondo del calcio, e tra le principali fonti al mondo per stabilire i valori di mercato di ogni singolo giocatore.



#	Giocatori	Naz	Età	Società	Valore di mercato
1	Victor Osimhen Punta centrale		24		120,00 mln €
2	Rafael Leão Ala sinistra		24		90,00 mln €
3	Lautaro Martínez Punta centrale		25		85,00 mln €
4	Khvicha Kvaratskhelia Ala sinistra		22		85,00 mln €
5	Nicolò Barella Centrale		26		75,00 mln €
6	Dusan Vlahovic Punta centrale		23		70,00 mln €
7	Alessandro Bastoni Difensore centrale		24		60,00 mln €
8	Theo Hernández Terzino sinistro		25		60,00 mln €
9	Min-jae Kim Difensore centrale		26		60,00 mln €

Figura 2: esempio di dataset di Transfermarkt dei valori di mercato della Serie A

FBREF: un sito che fa parte del "gruppo" Sport Reference, e che in particolare si occupa di riportare statistiche relative al mondo del calcio, sia a livello di competizioni, che a livello di società e addirittura singoli giocatori.

Rk	Squad	MP	W	D	L	GF	GA	GD	Pts	Pts/MP	xG	xGA	xGD	xGD/90	Attendance	Top Team Scorer	Goalkeeper
1	Barcelona	38	28	4	6	70	20	+50	88	2.32	75.5	33.2	+42.3	+1.11	83,498	Robert Lewandowski - 23	Marc-André ter Stegen
2	Real Madrid	38	24	6	8	75	36	+39	78	2.05	75.5	38.9	+36.6	+0.96	56,649	Karim Benzema - 19	Thibaut Courtois
3	Atlético Madrid	38	23	8	7	70	33	+37	77	2.03	62.3	41.3	+20.9	+0.55	55,800	Antoine Griezmann - 15	Jan Oblak
4	Real Sociedad	38	21	8	9	51	35	+16	71	1.87	52.7	33.4	+19.3	+0.51	32,189	Alexander Sørloth - 12	Álex Remiro
5	Villarreal	38	19	7	12	59	40	+19	64	1.68	61.7	50.3	+11.3	+0.30	16,406	Nicolas Jackson - 12	Pepe Reina
6	Betis	38	17	9	12	46	41	+5	60	1.58	51.8	50.1	+1.8	+0.05	49,635	Borja Iglesias - 15	Rui Silva
7	Osasuna	38	15	8	15	37	42	-5	53	1.39	39.4	45.7	-6.3	-0.17	19,655	Ante Budimir, Ezequiel Ávila - 8	Aitor Fernández
8	Athletic Club	38	14	9	15	47	43	+4	51	1.34	54.2	36.9	+17.3	+0.46	43,488	Iñaki Williams, Oihan Sancet - 10	Unai Simón
9	Mallorca	38	14	8	16	37	43	-6	50	1.32	35.2	46.4	-11.2	-0.30	14,799	Vedat Muriqi - 15	Predrag Rajković
10	Girona	38	13	10	15	58	55	+3	49	1.29	50.4	55.0	-4.6	-0.12	11,471	Valentín Castellanos - 13	Paulo Gazzaniga
11	Rayo Vallecano	38	13	10	15	45	53	-8	49	1.29	43.9	49.9	-6.0	-0.16	12,525	Isaac Palazón Camacho - 9	Stole Dimitrievski
12	Sevilla	38	13	10	15	47	54	-7	49	1.29	46.3	54.2	-7.9	-0.21	35,566	Youssef En-Nesyri - 8	Yassine Bounou
13	Celta Vigo	38	11	10	17	43	53	-10	43	1.13	42.8	45.5	-2.8	-0.07	14,425	Iago Aspas - 12	Agustín Marchesín, Iván Villar
14	Cádiz	38	10	12	16	30	53	-23	42	1.11	41.6	61.4	-19.9	-0.52	17,741	Theo Bongonda, Gonzalo Escalante - 4	Jeremías Ledesma
15	Getafe	38	10	12	16	34	45	-11	42	1.11	36.7	46.0	-9.4	-0.25	11,592	Enes Ünal - 14	David Soria
16	Valencia	38	11	9	18	42	45	-3	42	1.11	49.3	43.5	+5.8	+0.15	40,602	Samuel, Justin Kluivert - 6	Giorgi Mamardashvili
17	Almería	38	11	8	19	49	65	-16	41	1.08	45.5	62.5	-17.0	-0.45	12,961	El Bilal Touré - 7	Fernando
▼ 18	Valladolid	38	11	7	20	33	63	-30	40	1.05	38.6	64.9	-26.3	-0.69	20,943	Cyle Larin - 8	Jordi Masip
▼ 19	Espanyol	38	8	13	17	52	69	-17	37	0.97	48.3	59.8	-11.5	-0.30	21,601	Joselu - 16	Fernando Pacheco
▼ 20	Elche	38	5	10	23	30	67	-37	25	0.66	37.5	70.0	-32.6	-0.86	19,875	Lucas Boyé - 7	Édgar Badía

Figura 3: esempio di dataset di FBREF della Liga Spagnola, stagione 2022/23

Sono state usate due tecniche diverse per l'acquisizione dei dati da queste fonti: per Capology, è stato utilizzato il web scraping, grazie alla libreria *BeautifulSoup* presente in Python. Per FBREF e Transfermarkt, invece, è stata utilizzata una libreria di R ("*worldfootballR*") che permette l'acquisizione di dati da entrambi i siti.

In particolare, tramite tecniche di web scraping implementate con la libreria *BeautifulSoup*, sono stati acquisiti i dati relativi ai contratti dei singoli giocatori di otto competizioni differenti: la Premier League (la prima serie inglese), la Serie A (la prima serie italiana), la Primera Division (comunemente chiamata La Liga, la prima serie spagnola), la Bundesliga (la prima serie tedesca), la Ligue 1 (la prima serie francese), la Primeira Liga (comunemente nota come Liga Nos, la prima serie portoghese), la Major League Soccer (o abbreviata come MLS, la prima serie americana), e l'Eredivisie (la prima serie olandese).

Allo stesso modo, sono stati raccolti i dati di FBREF e di Transfermarkt attraverso diverse funzioni presenti all'interno della libreria *worldfootballR*. Per quanto riguarda Transfermarkt, sono state raccolte informazioni anagrafiche (come data di nascita, piede forte, altezza…), alcune informazioni relative al contratto (quando il giocatore si è unito alla squadra attuale, e il nome della squadra precedente), oltre ovviamente al valore di mercato. In particolare, quest'ultimo è frutto di una serie di valutazioni fatte dagli esperti del sito, che tengono conto di diversi fattori come le prestazioni, l'età, il ruolo, la reputazione del giocatore, l'interesse sul mercato per questo, eventuali trasferimenti avvenuti di recente…

Infine, sono state raccolte una serie di statistiche relative ai campionati della stagione 2022/23 dei calciatori delle 8 competizioni citate in precedenza. Sono state raccolte informazioni relative a diverse categorie, tra cui: statistiche generali, relative ai tiri, ai passaggi, alla fase difensiva, alla fase di possesso, al tempo di gioco…

I dati raccolti tramite web scraping dal sito Capology sono stati convertiti in formato *JSON*, con un file creato per ogni competizione e una chiave primaria per ogni giocatore che ha preso parte a quel campionato.

I dati di Transfermarkt raccolti tramite la libreria *worldfootballR* (e in particolare, con la funzione *tm_player_market_values(),* dove andava specificata la stagione e i campionati di interesse) venivano forniti come oggetto di tipo *dataframe*, avente 19 variabili e 5903 osservazioni. Inoltre, la funzione *player_dictionary_mapping()* permetteva di ottenere

un dataframe contenente i link di FBREF e di Transfermarkt di oltre 14000 calciatori, in modo tale che fosse resa più semplice l'unione dei dati importati da entrambi i siti.

Per importare i dati di FBREF, invece, è stata implementata una funzione che, dati il campionato, la stagione e il tier (ovvero il livello del campionato), creava e restituiva una lista, contenente un *dataframe* per ogni squadra presente nella competizione di interesse in quella stagione: in ognuno di questi dataframe erano raccolti i dati relativi a 9 aspetti dello sport (ovvero "standard", "shooting", "passing", "passing_types", "gca", "defense", "possession", "playing_time" e "misc") per ogni giocatore presente in ogni rosa di quel campionato. La funzione creata, dunque, estraeva il link del campionato di interesse (con la funzione *fb_league_urls()*), da cui venivano estratti i link di ogni singola squadra (con la funzione *fb_teams_urls()*), e successivamente estraeva i 9 *dataframe* relativi ai differenti aspetti del gioco per ogni singola squadra.

Successivamente, la funzione creata procedeva con due operazioni in parallelo: l'unione tramite dei *merge()* dei 9 dataframe trovati, per ogni squadra, e l'esclusione tramite la funzione *duplicated()* delle features che si ripetevano tra i differenti *dataframes* estratti. Infine, la funzione restituiva dunque la lista contenente i *dataframes* di ogni squadra completi di tutte le statistiche di interesse, estratte da FBREF, per ogni singolo giocatore.

Pulizia e integrazione dei dati

Una fase molto importante è quella della pulizia dei dati. In parte alcune operazioni sono già state svolte (come spiegato nella sezione relativa all'acquisizione dei dati), ma comunque altre operazioni sono state svolte successivamente all'acquisizione vera e propria dei dati.

Per quanto riguarda i dati estratti tramite web scraping con Capology, per il formato che ci interessava (ovvero un file *CSV* per ogni campionato, con tutti i dati dei giocatori della competizione), sono state fatte alcune operazioni sull'*htmltree* estratto tramite la libreria *BeautifulSoup*:

- è stato individuato l'elemento "var data", che contraddistingueva l'inizio dei dati presenti nella pagina html, ed individuato inizio e fine dei dati;
- inoltre, sono stati esclusi alcuni caratteri (in ordine "\n", " ", ",{" e la virgola finale), oltre ad aver diviso il tutto per ogni singolo giocatore con la funzione *split()*;
- successivamente, con una serie di *replace()* sono stati tolti una serie di caratteri presenti all'interno delle singole stringhe di ogni giocatore, in modo da mantenere solamente le variabili e i corrispondenti valori come coppie chiavi-valore leggibili e uniformi.
- infine, è stato creato il dizionario relativo al singolo campionato, in cui ad ogni chiave corrispondeva un singolo calciatore avente un dizionario con tutte le informazioni di interesse;
- in ultimo, il dizionario veniva salvato sia in formato *.json*, sia in formato *.csv* (quello utilizzato poi in fase di integrazione).

Per quanto riguarda i dati estratti da FBREF, gran parte delle operazioni di cleaning sono quelle riportate nella funzione da noi creata, denominata *stat_giocatori_campionato()*. Al completamento delle operazioni di estrazione dei dati e creazione delle liste di dataframe per ogni campionato, abbiamo creato altre due funzioni per il salvataggio dei dati in due formati diversi, *.csv* e *.json*:

- la funzione *json_convert_campionato()* creava una cartella per il campionato da salvare, e procedeva salvando ogni elemento (dunque ogni singolo dataframe) della lista in un file *.json* col nome della squadra corrispondente, ottenendo dunque una tabella contenente tutti i *.json* relativi a quel campionato contenenti le statistiche estratte dal sito FBREF.
- allo stesso modo, la funzione *csv_convert_campionato()* creava nella cartella creata in precedenza per il campionato ogni elemento (dunque ogni singolo dataframe) della lista in un file *.csv* col nome della squadra corrispondente.

La fase di integrazione dei differenti dataset è stata effettuata con Python, con una serie di *concat()* e di *merge()* effettuati nel seguente ordine:

- unione degli stipendi di Capology in un unico *dataframe* (denominato *stipendi*) che comprendesse tutti i dati degli stipendi di ogni campionato;
- unione dei valori di mercato, forniti da Transfermarkt in un unico *dataframe*, col *dataframe* ottenuto dagli stipendi di Capology (dal nome *stipendi_valoremercato*).

- unione del *dataframe* ottenuto al passaggio precedente col *dataframe* per il mapping degli URL di Transfermarkt e di FBREF (sovrascrivendo il *dataframe* denominato *stipendi_valoremercato*);
- left join di ogni *dataframe* contenuto nelle liste di *dataframe* dei singoli campionati, col *dataframe* ottenuto al passaggio precedente (ovvero *stipendi_valoremercato*): in questo modo, abbiamo ottenuto 8 *dataframes* (con nomi determinati dal campionato di interesse e la stagione, ad esempio "*seriea_2023*", ecc…).

Inoltre, prima di iniziare l'integrazione dei dati (con Python), abbiamo verificato ulteriormente le features presenti nei differenti *dataframes*, e alcune colonne si ripetevano; dunque, abbiamo rimosso le informazioni ridondanti.

Alla fine, ai dati di FBREF sono state aggiunte le seguenti informazioni, ottenute da Capology, Transkfermarkt e il *dataframe* per il mapping (il tutto contenuto nel dataframe *stipendi_valoremercato*):

- *"player_num"* per indicare il numero di maglia del calciatore;
- *"player_dob"* per la data di nascita del calciatore;
- *"player_height_mtrs"* per l'altezza, in metri, del calciatore;
- *"player_foot"* per il piede forte del calciatore;
- *"date_joined"* per il giorno in cui il giocatore si è unito alla squadra attuale;
- *"joined_from"* per la squadra da cui è arrivato il calciatore all'inizio del contratto in vigore;
- *"contract_expiry"* per la data di scadenza del contratto;
- *"player_market_value_euro"* per il valore di mercato del calciatore, in euro;
- *"player_url"* per l'url che riporta alla pagina web del giocatore su Transfermarkt.
- *"annual_gross_eur"*: lo stipendio lordo del calciatore in euro;
- *"annual_net_eur"*: lo stipendio al netto delle tasse del calciatore in euro;
- *"bonus_gross_eur"*: l'eventuale bonus presente nel contratto del calciatore;
- *"signed"*: la data in cui era stato firmato il contratto corrente (sia in caso di rinnovo, che in caso di arrivo nella squadra corrente);
- *"years"*: gli anni prima della scadenza del contratto corrente;
- *"total_gross_eur"*: quanto la società deve ancora pagare al giocatore per gli anni rimanenti di contratto, al lordo;
- *"total_net_eur"*: quanto la società deve ancora pagare al giocatore per gli anni rimanenti di contratto, al netto;
- *"release_eur"*: se presente nel contratto, l'eventuale valore della clausola rescissoria del calciatore, in euro;
- *"loan"*: la variabile che definisce se il giocatore è in prestito o no.
- *"UrlFBref"*: per l'url che riporta alla pagina web del giocatore su FBREF.

Inoltre, i dati relativi alle variabili "club" e "name" sono state rese minuscole per permettere il merging dei dataset di Capology con quelli di Transfermarkt e di FBREF.

Come risultato finale, dunque, abbiamo ottenuto delle liste di *dataframes* (una lista per ogni campionato, contenente un *dataframe* per ogni squadra). In totale abbiamo 8 liste di *dataframes* corrispondenti ai campionati di interesse, ognuna con un numero variabile di squadre (da 18 ad un massimo di 28 squadre per l'MLS), a loro volta con un numero variabile di giocatori per squadra (solitamente in un numero tra i 25 e i 40 per rosa), e con una serie di statistiche e dati relativi ai singoli giocatori che anche in questo caso variano in base al campionato e alla squadra di appartenenza: infatti alcune features di FBREF non sono state raccolte per alcuni campionati, ed altre sì.

I dati sono stati esportati in formato *.csv*, per essere inseriti nel database management system che abbiamo scelto in base alle caratteristiche dei nostri dati, agli obiettivi che ci siamo posti per il progetto e alle funzioni che garantisce il DBMS utilizzato.

Caricamento del dataset in MongoDB

Per la fase di data storage, abbiamo deciso di utilizzare **MongoDB**: la scelta è ricaduta su MongoDB in quanto si tratta di un document based management system in cui ogni documento dev'essere immagazzinato in una collezione.

La scelta di un database NoSQL risiede nel fatto che ci permette di superare alcuni limiti dei modelli relazionali (come ad esempio, la flessibilità di un modello NoSQL, molto utile nel nostro caso avendo un numero non predefinito di giocatori e di variabili che compongono ogni dataset dei vari campionati), e

poiché ci permette di eseguire alcune operazioni con grande rapidità (ad esempio, queries di ricerca nelle collezioni), nonché di aggiungere nuovi dati o di spostare un documento da una collezione all'altra in caso di trasferimento in campionati differenti. Ù

Inoltre, il fatto che sia "*schemaless*" permette di superare quello che sarebbe stato un problema se avessimo usato un RDBMS: per ogni giocatore abbiamo un numero differente di features, che può variare sia in base al campionato che alla squadra di appartenenza; un modello documentale ci permette di superare questo "limite".

Per il nostro caso, abbiamo pensato di creare una collezione per ogni campionato differente, e all'interno di ogni collezione abbiamo una serie di documenti che contengono i dati dei singoli giocatori, che compongono ogni rosa della competizione.

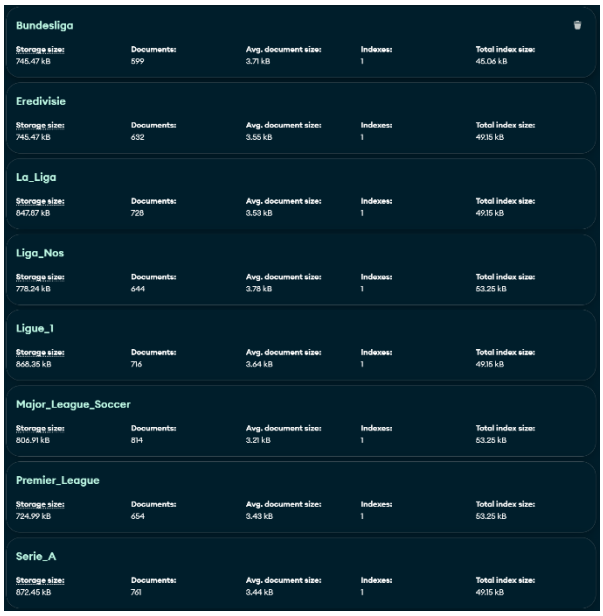


Figura 4: collezioni presenti nel database creato su MongoDB

In questo modo è possibile non solo eseguire query per documenti (dunque calciatori) all'interno di un campionato, ma anche di andare a trovare informazioni relative a squadre (tramite *aggregations* sfruttando la clausola *$group*), oltre a poter ricercare giocatori con determinate caratteristiche tra le differenti competizioni presenti nel dataframe o trovare informazioni relative ai giocatori e alle squadre in un particolare campionato.

Queries

Abbiamo provato ad eseguire alcune queries, provando ad immaginarci dirigenti di una squadra di calcio o procuratori sportivi, e a trovare soluzioni ad alcuni problemi pratici.

Ci siamo posti le seguenti domande:

1- Il Real Madrid, storico club spagnolo, ha perso la stella Karim Benzema, ed ora ha il compito non semplice di sostituirlo. Fortunatamente hanno un budget importante, e dunque ci siamo chiesti: quali sono dei nomi interessanti con cui pensare di sostituire il loro ormai ex centravanti?
2- Se un attaccante puntasse ad avere una nuova sfida personale per la sua carriera, e ad andare in una squadra nei top 5 campionati europei che non ha fatto bene dal punto di vista offensivo la scorsa stagione, quali sono delle possibili squadre che il suo procuratore gli potrebbe consigliare da risollevare dal punto di vista offensivo?

Per la prima domanda, abbiamo eseguito una serie di queries nelle pipeline *"aggregations"*: abbiamo innanzitutto cercato Karim Benzema, in modo tale da avere un'idea del suo apporto alla squadra dal punto di vista delle statistiche; successivamente abbiamo eseguito gli *$unionWith* con cui possiamo considerare tutte le collezioni (dunque, tutti i campionati), così da avere il dataset esteso, e abbiamo filtrato con *$match* cercando un attaccante che potesse avere caratteristiche simili a Benzema (dunque giocatori nel ruolo "FW", ovvero forward, con numeri simili o superiori di expected goals e expected assist). Il risultato è stato di 6 profili: Robert Lewandowski, Mohamed Salah, Jonathan David, Alexandre Lacazette, Kylian Mbappè e Mehdi Taremi sono i 6 nomi ottenuti per il post-Benzema al Real Madrid.

Per la seconda query, abbiamo eseguito sempre una pipeline *"aggregations"*: come nel caso di Benzema, abbiamo eseguito gli *$unionWith* con cui abbiamo potuto considerare tutte le competizioni, così da avere il dataset esteso; successivamente, abbiamo raggruppato tramite un *$group* i valori di expected goals per squadra, e infine ordinato con un *$sort* per ottenere i valori dal più basso al più alto. In questo modo, abbiamo ottenuto un documento per ogni squadra con i valori di expected goals totali ordinati dal basso verso l'alto: le prime squadre proposte sono la Sampdoria (34.5 expected goals), l'Augsburg (35.5 expected goals), il Mallorca (36.1 expected goals), il Lecce (36.3 expected goals) e l'Ajaccio (36.5 expected goals).

Data Quality

Per valutare la qualità dei dati abbiamo adottato la misura di *completeness* delle tabelle: essa rappresenta

un indicatore chiave per valutare il grado di completezza dei dati in un dataset. In particolare, indica la percentuale di data osservati come null presenti nel dataset.

Dai risultati ottenuti, emerge che il livello di *table completeness* delle tabelle è piuttosto basso per tutte le leghe prese in considerazione. Ad esempio, la Serie A presenta un livello di completezza pari al 0.209%, mentre la Bundesliga è al 0.148%. La Liga e l'Eredivisie raggiungono rispettivamente il 0.19% e il 0.178%. Analogamente, la Liga Nos e la Ligue 1 hanno livelli di completezza dell'0.13% e del 0.165%. La Premier League, nonostante sia leggermente più completa rispetto alle altre, registra comunque un livello di *table completeness* del 0.213%. Infine, a livello generale, tutto il dataset presenta un livello di *table completeness* pari al 0.187%.

Successivamente, è stata analizzata la consistency dei dati. La consistency si riferisce alla coerenza dei dati rispetto al loro significato e alle regole di validità. Una volta valutata la completezza delle tabelle delle diverse leghe di calcio, è stata condotta un'analisi per verificare se i dati fossero coerenti con le aspettative e rispettassero le regole di validità per ciascuna fonte di dati.

I risultati dell'analisi di consistency hanno rivelato che i dati presenti nel dataset finale erano coerenti con il loro significato e rispettavano le regole di validità definite per le diverse fonti di dati. Ciò indica che i dati raccolti da ciascuna lega di calcio erano conformi alle aspettative e non presentavano discrepanze o ambiguità significative che potessero compromettere l'interpretazione corretta dei dati stessi.

Spunti possibili per futuri miglioramenti

Il nostro progetto si presenta come una proposta di creare una banca dati di informazioni relative ai calciatori dei principali campionati mondiali, e ovviamente rispetto ai dati in possesso delle società professionistiche il nostro dataset ne contiene un numero limitato.

Tuttavia, anche con gli strumenti a nostra disposizione, un progetto come il nostro è migliorabile ulteriormente, in diversi modi:

- È possibile ampliare ulteriormente i dati presenti all'interno dei differenti dataset aggiungendo non solo i dati relativi alle competizione europee (Champions League, Europa League e Conference League), ma anche i dati relativi alle nazionali di calcio.
- Inoltre, è possibile ampliare i dataset inserendo anche le informazioni relative alle stagioni precedenti all'ultima, così da avere un quadro più completo dei differenti calciatori a disposizione (ad esempio, se un giocatore è rimasto fermo a lungo per un serio infortunio nella stagione appena terminata, le statistiche di questo ne hanno risentito in maniera importante).
- In aggiunta a quanto appena detto, è possibile ampliare ulteriormente la banca dati aggiungendo ulteriori campionati (anche se sui siti utilizzati da noi, altri campionati non avevano la stessa quantità di dati presenti per quelli riportati, ma solamente una piccola parte).
- È possibile automatizzare il dataset in modo da inserire le informazioni aggiornate, di giornata in giornata, durante la stagione sportiva. In questo modo, ci si potrebbe muovere tempestivamente, anticipando la concorrenza su alcuni prospetti interessanti del panorama calcistico mondiale, oppure si potrebbero trovare insight interessanti da sfruttare nella preparazione delle partite, in modo da poter sfruttare i propri punti di forza e le debolezze delle altre squadre.

Questi sono solo alcuni degli spunti possibili su cui sviluppare ulteriormente il progetto da noi proposto con questo lavoro. Il calcio sembra andare sempre più in questa direzione, dove i dati hanno sempre più un ruolo fondamentale in moltissimi aspetti dello sport.

Fonti

https://fbref.com/en/

https://www.transfermarkt.it/

https://www.capology.com/

https://jaseziv.github.io/worldfootballR/index.html

https://www.mongodb.com/