ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

SCUOLA DI SCIENZE

Dipartimento di Informatica – Scienza e Ingegneria

Corso di Laurea Triennale in Informatica per il Management

Il concetto di Extended Goals nella Football Data Analysis: uno studio comparativo relativo a Uefa Euro 2020

Relatore:
Prof. Marco Roccetti

Presentata da:
Filippo Berveglieri

Sessione IV

Anno Accademico 2022-2023

Indice

Introduzione

Dagli anni 2000 in poi l'utilizzo della statistica e del dato è divenuto sempre più parte integrante dello sport, con il calcio a rappresentare una locomotiva trainante, grazie anche all'ampia scala di risorse di cui dispone [1].
Il calcio moderno, profondamente supportato dall'evoluzione tecnologica, è sempre più influenzato dalla crescente importanza attribuita all'analisi dei dati [2]. Società e staff tecnici continuano ad integrare e potenziare l'analisi del dato in molte delle scelte da prendere, essa sia nella fase di scouting, piuttosto che nella gestione e prevenzione di infortuni o una semplice strategia di squadra, mentre la maggior parte dei tifosi vede questo fenomeno come il tentativo di razionalizzare il calcio, da sempre visto come materia emotiva e irrazionale [3].

In questo contesto di trasformazione, l'obiettivo principale di questa tesi è esaminare il ruolo dei dati nel calcio, approfondendo la comprensione di come i Big Data siano integrati nello sport. In particolare, si punta a esplorare le Key Performance Indicators (KPIs), focalizzando l'attenzione sugli Expected Goals (xG). Attraverso l'applicazione di modelli statistici, l'intento è fornire un contributo scientifico sul valore dell'xG e sul rapporto tra esso e i gol effettivamente realizzati.

Per raggiungere questo obiettivo, è stata condotta una fase iniziale di studio del contesto, osservando il progresso dall'introduzione professionale dell'analisi dei dati fino all'attuale ubiquità del data analyst all'interno degli staff delle squadre. Successivamente, è stato identificato il dataset più idoneo per il quesito iniziale, optando per StatsBomb, rivelatosi completo e organizzato per quanto riguarda gli xG. Nella fase successiva, sfruttando StatsBomb e programmando in Python e RStudio, sono stati estratti e organizzati i dati necessari per procedere

all'applicazione dei modelli statistici, con particolare attenzione al test di Kolmogorov-Smirnov.

La tesi si sviluppa in quattro capitoli distinti. Il primo capitolo è un'introduzione mirata a chiarire il contesto, il secondo affronta il tema dei Big Data e l'utilizzo di StatsBomb, mentre il terzo e il quarto sono dedicati rispettivamente ai test statistici e alle conclusioni che possono essere derivate dall'analisi compiuta.

Capitolo 1

Football Data Analysis: i Big Data e l'analisi di dati nel calcio

1.1: L'utilizzo crescente della statistica nel calcio

In questo paragrafo si osserva come la statistica è entrata nel mondo del calcio con l'ausilio di cenni storici e attraverso il paragone con la pallacanestro. Verranno anche accennate alcune tecnologie che ancora oggi vengono sfruttate e, infine, si osserverà come le strutture societarie e gli staff tecnici e medici hanno cambiato i loro assetti nel corso degli anni

1.1.1: Evoluzione dell'impiego del dato da parte dei professionisti

L'analisi statistica applicata al calcio sta attraversando un periodo di forte espansione, particolarmente evidente nell'ultimo decennio, che ha posizionato questo sport al vertice per la produzione di dati, superando persino la pallacanestro.

La pallacanestro, grazie alla sua struttura di gioco, ha sempre suscitato studi statistici approfonditi, offrendo spettatori e professionisti una conoscenza dettagliata dello sport[1]. È noto, infatti, che negli ultimi anni, durante una partita di pallacanestro, siamo in grado di conoscere in tempo reale una serie di statistiche relative a ogni giocatore, tra cui la percentuale di successo nei tiri da 2 e 3 punti, i rimbalzi offensivi e difensivi, e i possessi persi. In questo contesto, si è passati da affermazioni generiche come "Lebron James tira con il 50% da 3 punti" a dichiarazioni più dettagliate,

[1] https://www.ilsole24ore.com/art/data-analytics-sport-mercato-3-miliardi-dollari-AEmxJepB

che includono informazioni specifiche come il luogo in cui effettua i tiri e la situazione di gioco[2].

D'altra parte, nel calcio, la statistica ha occupato a lungo un ruolo marginale, relegato a un'analisi post-partita con dati rudimentali riguardanti il possesso palla, i tiri in porta, i passaggi e le ammonizioni. Un esempio concreto è rappresentato dalla finale del mondiale 2006 tra Italia e Francia, per la quale, a quel tempo, era difficile reperire statistiche più dettagliate rispetto a un semplice tabellino della partita.

Negli anni seguenti sono state introdotte statistiche leggermente più sofisticate riguardanti gli eventi del match per poi arrivare alla prima vera innovazione: i sistemi di tracciamento. I sistemi di tracciamento, detti anche GPS, stanno vivendo una continua evoluzione dentro ad un mondo già di per sé in evoluzione. I primi GPS, che per i tempi rappresentavano già una scoperta, erano strumenti che consentivano di controllare il chilometraggio del calciatore, la velocità media, la velocità massima e poco altro. Oggi, gli strumenti che i calciatori indossano quando giocano, sono strumenti potentissimi in grado di combinare dati posizionali e dati medici che non si limitano più a produrre output basici ma che iniziano a fornire informazioni quali posizionamento del calciatore, VO2 massima e media, consumi in termini calorici e tanti altri dati di carattere medico[3].

Con l'introduzione dei sistemi di tracciamento, sono emerse le heatmap, ovvero mappe di calore che rappresentano le zone di maggior attività di un

[2] https://www.rainews.it/articoli/2023/05/pallacanestro-basket-olimpia-ettore-messina-gregg-popovich-san-antonio-spurs-big-data-statistica-cdcdb120-1d3e-4c83-8557-5ecb49aa953a.html
[3] https://www.hitrech.com/spogliatoio/il-gps-nel-calcio-e-davvero-la-migliore-tecnologia.html

calciatore durante una partita. Queste mappe forniscono spunti preziosi, come le zone in cui un giocatore ha toccato più palloni o quelle in cui ha corso maggiormente[4]. Ad esempio, possedere le heatmap dei calciatori avversari può aiutare a identificare chi è coinvolto maggiormente nelle fasi di impostazione, sviluppo e rifinitura.

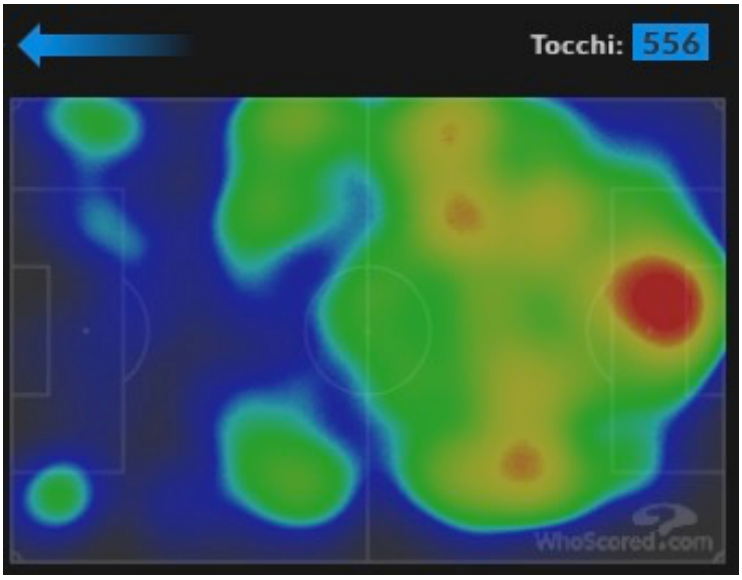


Figura 1: Tocchi palla dell'Inter nella partita Inter-Fiorentina (28-01-2024)[5]

Come è facile intuire, partendo da questi tipi di dati, quali le heatmap avversarie, si iniziano a sviluppare le prime strategie tecnico-tattiche per affrontare la partita. In questo periodo storico, gli addetti ai lavori cominciano a depositare fiducia nella statistica applicata al calcio, dando vita alla figura ampiamente dibattuta del Match Analyst.

[4] https://medium.com/calcio-datato/quali-sono-le-zone-di-campo-pi%C3%B9-battute-dalle-squadre-di-serie-a-3c88c76625b7
[5] https://it.whoscored.com/Matches/1746244/Live/Italia-Serie-A-2023-2024-Fiorentina-Inter

Il Match Analyst è una figura che, ancora oggi, presenta notevoli differenze a seconda del livello della società calcistica. Nel contesto dilettantistico, se presente, il Match Analyst può essere una figura semplice, dedicata all'analisi delle registrazioni della partita avversaria. Al contrario, ai livelli professionistici più elevati, il Match Analyst diventa un elemento chiave dello staff tecnico. Ma questo verrà approfondito nel paragrafo successivo.

A questo punto, grazie anche a questa nuova figura, la consapevolezza degli staff e delle società è cresciuta, iniziando a percepire la necessità di un maggior numero di dati da analizzare[6]. Negli ultimi anni, le società specializzate nell'elaborazione dati hanno fatto il loro ingresso deciso nel mondo del calcio. Queste aziende si dedicano alla creazione di database ricchissimi, che vengono successivamente venduti alle sempre più numerose società professionistiche interessate all'analisi avanzata. Da questa evoluzione è nata la figura del Football Data Analyst, il cui compito principale è estrarre informazioni di vario genere (performance, scouting, spunti tattici, ecc.) da tali database[7].

Attualmente, il Football Data Analyst è una figura in evoluzione e trova la sua collocazione principalmente nelle società di primissimo livello. Tuttavia, ci sono eccezioni come il Manchester City, noto per le sue ampie risorse finanziarie, che ha recentemente affidato la gestione completa dello storage e della protezione dei dati ad Acronis[8][9], una multinazionale specializzata.

[6] https://www.wired.it/scienza/lab/2016/03/25/calcio-big-data/
[7] https://wylab.net/formazione/football-data-analyst/
[8] https://www.mancity.com/club/partners/acronis
[9] https://www.acronis.com/en-eu/sports/manchester-city/

Con l'ascesa dei Big Data, il calcio è diventato lo sport con la mole di dati più grande prodotta, superando addirittura la pallacanestro, nonostante la struttura del gioco sembrasse originariamente renderlo meno incline a questo tipo di analisi. Mentre nella pallacanestro si verificano numerosi eventi in ogni partita, nel calcio le partite spesso terminano con punteggi più bassi e con un numero di azioni limitato. Tutto ciò è sintetizzato nel concetto di "riproducibilità dell'evento", che costituisce un elemento chiave nella statistica.

Oggi, il calcio è riuscito a integrare appieno il mondo della statistica grazie alla registrazione dettagliata di un numero crescente di eventi e alla creazione di Key Performance Indicators (KPIs), ovvero indicatori di performance. Questo approccio consente di formulare affermazioni specifiche, come ad esempio nel caso del Real Madrid: "Ogni 3 dribbling di Vinicius negli ultimi 30 metri, c'è un gol."

1.1.2: Il mutamento degli staff con l'avvento della statistica

L'introduzione della statistica nel contesto calcistico ha innescato un profondo mutamento nella struttura e nella composizione degli staff tecnici delle squadre, ridefinendo radicalmente il modo in cui vengono prese le decisioni e condotte le attività quotidiane. Per comprendere appieno questo cambiamento, è essenziale esplorare l'evoluzione storica degli staff tecnici e delineare il passaggio da un modello tradizionale a uno più orientato all'analisi dati.

In passato, gli staff tecnici delle squadre di calcio erano prevalentemente costituiti da figure come allenatori, preparatori atletici e osservatori di

talenti. Questi professionisti, dotati di esperienza sul campo e spesso influenzati dalla loro carriera da giocatori, basavano le proprie decisioni principalmente su intuizioni, conoscenza del gioco e apprezzamento dell'abilità individuale. Le strategie di gioco erano spesso sviluppate in base a metodi tradizionali, e le osservazioni durante le partite dipendevano principalmente dall'occhio umano e dall'esperienza tattica dei membri dello staff. Con l'avvento della statistica applicata al calcio, la dinamica degli staff tecnici ha subito una trasformazione significativa. Nuove figure specializzate hanno fatto la loro comparsa, tra cui i Match Analysts e i Football Data Analysts. Questi professionisti hanno introdotto un approccio più razionale e basato sui dati nell'analisi delle partite e nella formulazione delle strategie di gioco.

Anche nei reparti medici e di preparazione atletica, la competenza analitica è diventata fondamentale. I preparatori atletici basano ora i loro programmi di allenamento su dati fisiologici e biomeccanici, monitorati attraverso avanzati dispositivi tecnologici. I medici della squadra integrano le informazioni statistiche nella gestione della salute degli atleti, prevenendo infortuni e ottimizzando le prestazioni.

Persino nel processo di scouting, tradizionalmente guidato dall'esperienza umana, gli algoritmi e l'analisi statistica giocano un ruolo crescente. Le squadre utilizzano modelli predittivi per identificare giocatori con determinate caratteristiche che si adattano al proprio stile di gioco, migliorando così l'efficienza nella ricerca di talenti.

In conclusione, il passaggio da staff tecnici tradizionali a quelli orientati all'analisi dati rappresenta un capitolo significativo nella storia del calcio

moderno. Questa evoluzione ha coinvolto non solo gli aspetti tecnici, ma ha permeato tutti gli ambiti dello sport, evidenziando l'importanza cruciale dell'approccio basato sui dati in ogni fase della gestione calcistica contemporanea.

Questo per sottolineare quanto è cambiata la mentalità e l'approccio verso qualsiasi processo all'interno delle società di calcio.

1.2: Le Key Performance Indicators (KPI) nel calcio

I KPI, comunemente impiegati nel contesto aziendale e nella gestione aziendale, rappresentano indicatori utilizzati per valutare le performance di un'azienda, focalizzandosi sulle operazioni specifiche[10].

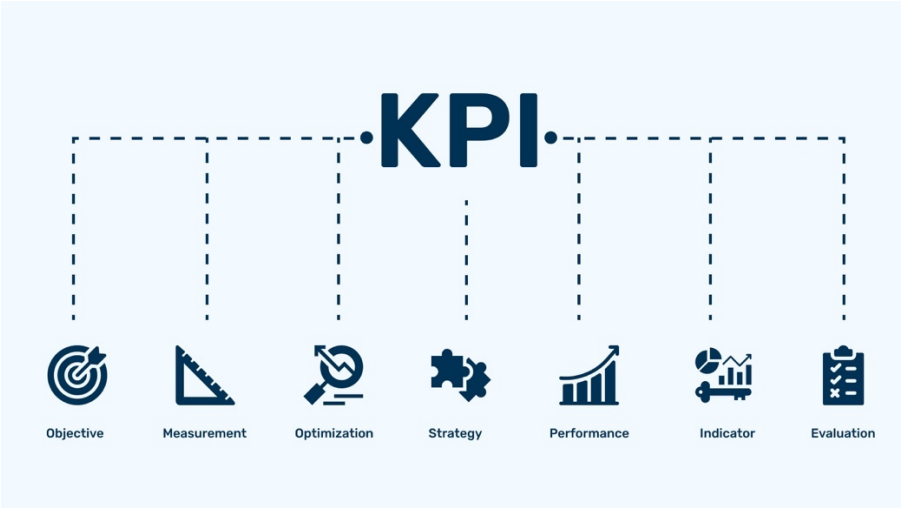


Figura 2: Key Performance Indicators

[10] https://strategiedigitali.net/kpi-aziendali-definizione-significato-ed-esempi/

Nel contesto calcistico, il concetto di KPI è analogo, poiché queste metriche sono essenziali per valutare le prestazioni di un calciatore. Naturalmente, a seconda del ruolo del calciatore, vengono analizzati indicatori differenti. Nell'analisi delle prestazioni di un attaccante, ad esempio, si considerano gli Expected Goals (xG) confrontati con i gol effettivamente realizzati, piuttosto che l'indice di coinvolgimento con la squadra. Invece, in un difensore centrale, si dà maggior rilevanza a indicatori come la percentuale di duelli aerei vinti.

In questo capitolo, mediante l'osservazione di alcuni grafici, esamineremo alcune delle KPI più significative nel contesto calcistico.

1.2.1: Expected Goals (xG)

Il parametro più utilizzato e conosciuto tra quelli menzionati, nonché l'argomento principale di questa tesi, è l'Expected Goals (xG). Esso consiste in un valore assegnato al tiro prima che questo venga effettuato, rappresentando la probabilità che tale tiro si trasformi in gol[11]. Questo valore varia da 0 a 1, e maggiore è la sua prossimità all'1, maggiore è la probabilità di segnare un gol. Ad esempio, un calcio di rigore ha un xG pari a 0,78, indicando che ha il 78% di probabilità di essere convertito in golL'assegnazione del valore ad un tiro avviene tramite un modello statistico ben preciso, legato ad una analisi di ben 300mila tiri[12]. I fattori che influenzano il valore finale sono molteplici, ma tra quelli che incidono più significativamente vi sono la posizione di tiro, il tipo di tiro, il contesto dell'azione e la presenza di disturbo avversario.

[11] https://www.ultimouomo.com/expected-goals/
[12] https://www.90min.com/it/posts/cosa-sono-e-come-calcolare-gli-expected-goal-xg

Come evidenziato nel boxplot sottostante, elaborato su dati di StatsBomb, la maggior parte dei gol è stata realizzata con tiri da posizioni ravvicinate.

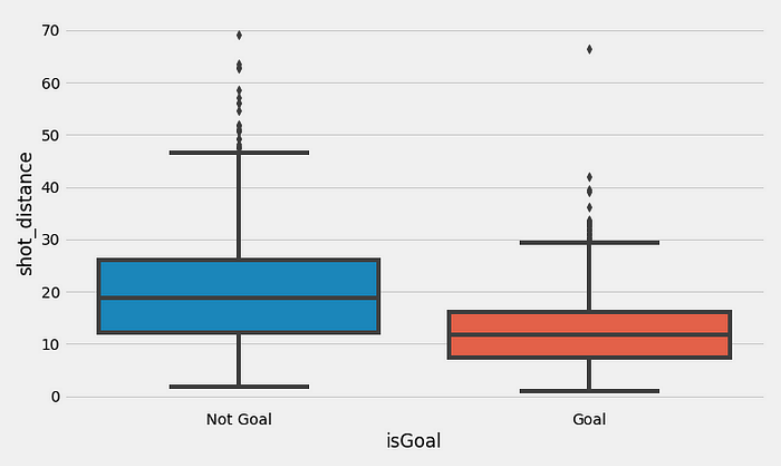


Figura 3: boxplot sulle distanze dei tiri

Un ulteriore elemento determinante è l'angolo di tiro, come illustrato nel boxplot successivo. I tiri effettuati da una posizione più centrale mostrano una percentuale di realizzazione maggiore.

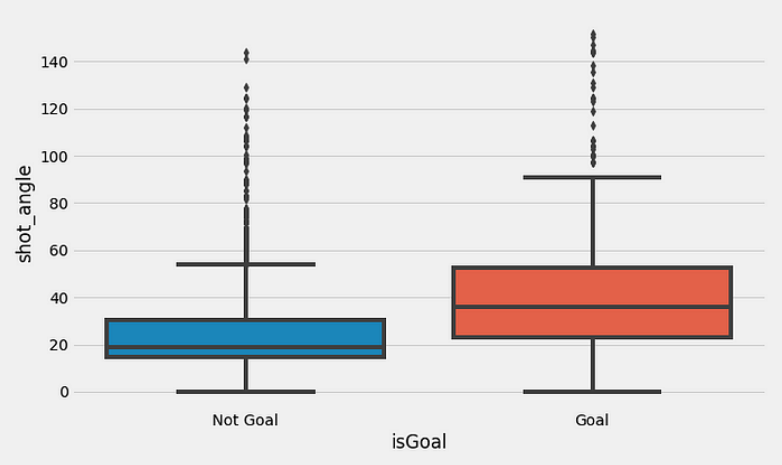


Figura 4: boxplot sull'angolo di tiro

L'Expected Goals risulta particolarmente utile, tra le varie applicazioni, per valutare se una squadra si trova in un momento di overperformance o

underperformance. Se una squadra segna più gol rispetto a quanto suggerito dall'xG, significa che sta overperformando; viceversa, se segna meno, sta underperformando.

1.2.2: Expected Goals on Target (xGoT)

Un altro KPI altrettanto significativo è l'xGoT. Mentre nell'xG il valore viene assegnato prima che il tiro sia effettuato, nell'xGoT viene assegnato dopo il suo scocco, considerando così la traiettoria effettiva della palla[13]. A differenza dell'xG, l'xGoT tiene conto della destinazione finale del tiro, quindi varia a seconda se il tiro è centrale o termina all'incrocio dei pali. Ad esempio, l'xGoT di un tiro all'incrocio dei pali sarà molto più elevato rispetto allo stesso tiro centrato.

Questo indice risulta prezioso per valutare l'abilità del tiratore. Se l'xGoT è superiore all'xG, significa che il calciatore ha effettuato un tiro di qualità; al contrario, se riduce l'xG, indica che il tiro è stato al di sotto della media. Tale affermazione si basa sul fatto che nell'Expected Goal è considerato un tiro con un piazzamento medio.

Un esempio concreto può essere illustrato attraverso due immagini relative a un tiro di Candreva, fornite dalla Lega Serie A[14].

Prima del suo scocco, il tiro aveva una percentuale di realizzazione del 2%, giustificata dai fattori elencati nell'immagine.

[13] https://www.eurosport.it/calcio/cosa-sono-gli-expected-goals-e-a-cosa-servono-le-5-statistiche-avanzate-piu-utilizzate-per-capire-me_sto9204614/story.shtml
[14] https://thegegenpress.it/glossario-statistico/

Figura 5: xG tiro di Candreva

L'xGoT, invece, considerando che il pallone termina sotto l'incrocio dei pali, gli assegna una probabilità di realizzazione del 60%.

Figura 6: xGoT del tiro di Candreva

1.2.3: Expected Assists (xA)

Un altro indicatore cruciale nell'ambito dell'analisi delle prestazioni calcistiche è l'Expected Assists (xA). Questo parametro si concentra sull'aspetto creativo dei giocatori, valutando la probabilità che un

determinato passaggio porti a un assist[15]. Così come nell'xG, l'assegnazione del valore avviene attraverso un modello statistico, considerando vari fattori come la posizione del passaggio, la distanza percorsa dalla palla e altri elementi contestuali.

L'xA risulta particolarmente utile per valutare l'efficacia dei giocatori nel fornire opportunità di gol ai compagni. Un valore elevato indica che il giocatore ha effettuato passaggi che, statisticamente, avevano buone probabilità di portare a un assist.

[15] https://www.statsperform.com/resource/expected-assists-in-context-2/

Capitolo 2

Estrazione e preparazione dei dati

Per condurre un'analisi dati di qualità, è imperativo disporre di un dataset che contenga le informazioni pertinenti. Nel contesto dell'analisi dei dati nel calcio, esistono numerosi dataset, ciascuno concepito per specifici scopi d'uso. Un dataset impiegato dal match analyst per studiare l'organizzazione difensiva della squadra avversaria sarà inevitabilmente diverso da quello utilizzato dal direttore sportivo per valutare l'acquisto di un nuovo attaccante. In questo capitolo, esploreremo le ragioni alla base della scelta di utilizzare StatsBomb come fonte principale di dati e successivamente analizzeremo come, mediante l'uso di Python, RStudio e Jupyter Notebook, ho estratto e preparato i dati, rendendoli pronti per l'applicazione dei modelli statistici.

2.1: StatsBomb

StatsBomb è un'azienda relativamente giovane, fondata da analisti per analisti, con una notevole fiducia nelle proprie competenze. Sul loro sito, si autodefiniscono i leader del settore e affermano di aver rivoluzionato la comprensione del calcio. L'azienda fornisce servizi non solo ai professionisti del calcio, raccogliendo dati tecnici per le prestazioni, lo scouting e l'analisi avversaria, ma si rivolge anche a media e organizzazioni di scommesse, collaborando nella creazione di modelli predittivi avanzati.

Tra i professionisti del settore, StatsBomb è apprezzata per la vastità dei dati raccolti all'interno di competizioni specifiche e per la loro ampia copertura di eventi. Con 3400 eventi registrati per partita e la copertura di 120 competizioni[16], l'azienda offre un ampio spettro di informazioni. Inoltre, StatsBomb si è specializzata nel campo della scoperta di talenti, offrendo strumenti dedicati.

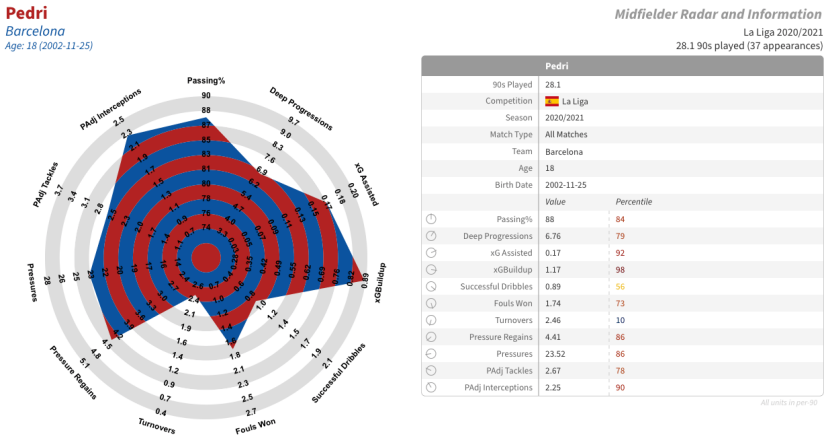


Figura 7: Analisi di Pedri con StatsBomb IQ

Riguardo alla mia scelta di utilizzare StatsBomb, ho iniziato a esplorare il mondo della Football Data Analysis alcuni mesi fa e, per fare ciò, ho cercato un dataset gratuito che mi permettesse di condurre esperimenti. In questo contesto, StatsBomb si è rivelata una scelta eccezionale, offrendo un dataset con un'ampia copertura di competizioni e ricco di eventi per ogni partita. Inizialmente, ho effettuato esperimenti semplici, come la creazione di mappe dei tiri di un calciatore o dei passaggi di una squadra. Successivamente, ho combinato diversi tipi di dati, arrivando infine al mio progetto di tesi, per il quale ho utilizzato prevalentemente i dati forniti da StatsBomb.

[16] https://statsbomb.com/what-we-do/soccer-data/360-2/

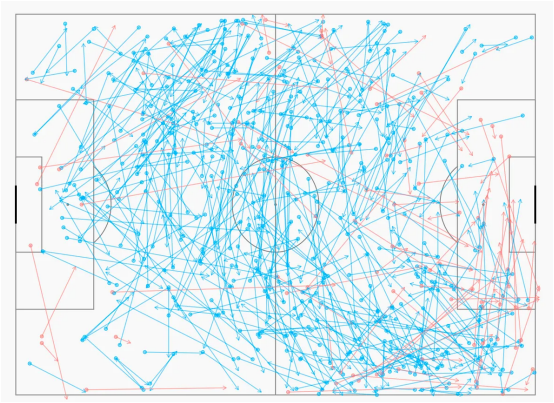


Figura 8: Passaggi dell'Arsenal WFC in un match di campionato

Un elemento determinante nella mia scelta è stato anche il fatto che StatsBomb fornisce librerie e API compatibili con linguaggi di programmazione come R e Python, che conoscevo già.

2.1.1: Accesso ai dati tramite API

Le API, acronimo di "Interfacce di Programmazione Applicativa" (Application Programming Interfaces), sono protocolli e strumenti che permettono a diversi software di comunicare tra loro[17]. In termini più semplici, sono set di regole che consentono a programmi diversi di scambiare informazioni tra di loro in modo strutturato e standardizzato.

Volgarmente, un'API come una sorta di "ponte" virtuale tra due applicazioni. Un'applicazione può richiedere l'accesso a determinate funzionalità o dati di un'altra applicazione attraverso l'API, senza dover conoscere i dettagli interni del funzionamento di quest'ultima. Questo facilita l'integrazione di diverse applicazioni e consente loro di collaborare senza la necessità di rivelare il proprio codice sorgente.

[17] https://www.ibm.com/it-it/topics/api

Nel contesto di StatsBomb e della Football Data Analysis, le API offrono un mezzo attraverso il quale possiamo accedere in modo strutturato alle informazioni dettagliate presenti nel loro dataset. Attraverso le API, possiamo effettuare richieste specifiche per ottenere dati su eventi di gioco, statistiche dei giocatori e altro ancora. La loro compatibilità con linguaggi di programmazione comuni come R e Python semplifica l'elaborazione e l'analisi dei dati nel nostro progetto.

2.1.2: Struttura dei dati StastBomb

Come precedentemente menzionato, StatsBomb offre una vasta gamma di competizioni nella sua versione gratuita. I dati possono essere consultati in formato JSON per una visualizzazione diretta, o, alternativamente, possono essere elaborati in ambienti come Python attraverso Jupyter Notebook, offrendo la possibilità di eseguire operazioni sui dati.

Per ottenere il dataset in formato JSON, è sufficiente recarsi su GitHub e avviare il download. Per lavorare con i dati in Jupyter Notebook, è necessario importare il modulo sb dalla libreria statsbombpy.

Nel contesto di Jupyter, le competizioni sono organizzate con ciascuna competizione rappresentata come una riga e gli attributi associati presentati nelle colonne.

```
In [3]: competitions = sb.competitions()
        pd.set_option('display.max_rows', None)
        competitions
```

ompetition_id	season_id	country_name	competition_name	competition_gender	competition_youth	competition_international	season_name	match_updated
9	27	Germany	1. Bundesliga	male	False	False	2015/2016	2023-12-12T07:43:33.436182
16	4	Europe	Champions League	male	False	False	2018/2019	2023-03-07T12:20:48.118250
16	1	Europe	Champions League	male	False	False	2017/2018	2021-08-27T11:26:39.802832
16	2	Europe	Champions League	male	False	False	2016/2017	2021-08-27T11:26:39.802832
16	27	Europe	Champions League	male	False	False	2015/2016	2021-08-27T11:26:39.802832
16	26	Europe	Champions League	male	False	False	2014/2015	2021-08-27T11:26:39.802832
16	25	Europe	Champions League	male	False	False	2013/2014	2021-08-27T11:26:39.802832
16	24	Europe	Champions League	male	False	False	2012/2013	2021-08-

Figura 9: Visualizzazione competitions Jupyter Notebook

Come evidenziato, StatsBomb non fornisce informazioni dettagliate sulle competizioni, come la nazione ospitante, lo stadio della finale e la squadra vincente, poiché si concentra principalmente sulla raccolta degli eventi relativi alle partite. La visualizzazione in formato JSON presenta tutte le coppie chiave-valore, con la maggior parte degli attributi che riflettono la disponibilità e l'aggiornamento dei dati, mentre solo alcuni forniscono informazioni sulla competizione.


```json
[ {
  "competition_id" : 9,
  "season_id" : 27,
  "country_name" : "Germany",
  "competition_name" : "1. Bundesliga",
  "competition_gender" : "male",
  "competition_youth" : false,
  "competition_international" : false,
  "season_name" : "2015/2016",
  "match_updated" : "2023-12-12T07:43:33.436182",
  "match_updated_360" : null,
  "match_available_360" : null,
  "match_available" : "2023-12-12T07:43:33.436182"
}, {
```


Figura 10: Visualizzazione json di competitions

Due attributi cruciali sono *competition_id* e *season_id*. Ogni competizione e ogni stagione hanno il proprio ID, e per ottenere informazioni sui match di una specifica competizione è necessario chiamare la funzione *matches* fornendo gli ID di competizione e stagione come attributi. Ad esempio, se l'ID della competizione FIFA World Cup è 43 e quello della stagione 2022 è 106, la chiamata sarebbe *sb.matches(competition_id=43, season_id=106)*.

La visualizzazione delle partite di una competizione mostra un elenco di partite con attributi alla partita stessa. Per esempio:

- *match_id*: è l'attributo principale che ci consente di poter consultare gli eventi relativi a quella partita
- *match_date* e *kick_off:* indicano data e orario della partita
- *home_team* e *away_team*: indicano le due squadre
- *competition_stage:* indica il turno della competizione in cui si è svolta quella partita (gironi, ottavi di finale, quarti di finale,…)
- *referee*: nome dell'arbitro
- *home_managers* e *away_managers:* sono i nomi dei due allenatori

	match_id	match_date	kick_off	competition	season	home_team	away_team	home_score	away_score	match_status
0	3857256	2022-12-02	21:00:00.000	International - FIFA World Cup	2022	Serbia	Switzerland	2	3	available
1	3869151	2022-12-03	21:00:00.000	International - FIFA World Cup	2022	Argentina	Australia	2	1	available
2	3857257	2022-11-30	17:00:00.000	International - FIFA World Cup	2022	Australia	Denmark	1	0	available
3	3857258	2022-11-24	21:00:00.000	International - FIFA World Cup	2022	Brazil	Serbia	2	0	available
4	3857288	2022-11-26	12:00:00.000	International - FIFA World Cup	2022	Tunisia	Australia	0	1	available

Figura 11: Visualizzazione delle partite di una competizione

Infine, selezionando una specifica partita, è possibile esaminare tutti gli eventi relativi ad essa, che costituiscono il focus principale della raccolta dati di StatsBomb.

```
Index(['50_50', 'bad_behaviour_card', 'ball_receipt_outcome',
       'ball_recovery_offensive', 'ball_recovery_recovery_failure',
       'block_deflection', 'block_offensive', 'carry_end_location',
       'clearance_aerial_won', 'clearance_body_part', 'clearance_head',
       'clearance_left_foot', 'clearance_other', 'clearance_right_foot',
       'counterpress', 'dribble_nutmeg', 'dribble_outcome', 'dribble_overrun',
       'duel_outcome', 'duel_type', 'duration', 'foul_committed_advantage',
       'foul_committed_card', 'foul_committed_offensive',
       'foul_committed_penalty', 'foul_committed_type', 'foul_won_advantage',
       'foul_won_defensive', 'foul_won_penalty', 'goalkeeper_body_part',
       'goalkeeper_end_location', 'goalkeeper_outcome', 'goalkeeper_position',
       'goalkeeper_technique', 'goalkeeper_type', 'id', 'index',
       'interception_outcome', 'location', 'match_id', 'minute', 'off_camera',
       'out', 'pass_aerial_won', 'pass_angle', 'pass_assisted_shot_id',
       'pass_body_part', 'pass_cross', 'pass_deflected', 'pass_end_location',
       'pass_goal_assist', 'pass_height', 'pass_inswinging', 'pass_length',
       'pass_outcome', 'pass_outswinging', 'pass_recipient',
       'pass_shot_assist', 'pass_switch', 'pass_technique',
       'pass_through_ball', 'pass_type', 'period', 'play_pattern', 'player',
       'player_id', 'position', 'possession', 'possession_team',
       'possession_team_id', 'related_events', 'second', 'shot_aerial_won',
       'shot_body_part', 'shot_end_location', 'shot_first_time',
       'shot_freeze_frame', 'shot_key_pass_id', 'shot_one_on_one',
       'shot_outcome', 'shot_statsbomb_xg', 'shot_technique', 'shot_type',
       'substitution_outcome', 'substitution_replacement', 'tactics', 'team',
       'team_id', 'timestamp', 'type', 'under_pressure'],
      dtype='object')
```

Figura 12: eventi raccolti da StatsBomb per ogni match

È importante notare che oltre agli eventi tradizionali come tiri, passaggi e falli, StatsBomb fornisce dati dettagliati come *position*, indicante la posizione in cui si verifica l'evento, tipo di fallo commesso e *shot_statsbomb_xg*, che sarà un elemento chiave in questa tesi.

In questo paragrafo, ho fornito una panoramica generale della struttura del dataset, mentre nei paragrafi successivi inizierò a parlare dell'estrazione specifica dei dati per il mio progetto.

2.1: Estrazione indici xG delle partite di UEFA EURO 2020

Prima di approfondire il processo di estrazione dati, è essenziale definire chiaramente l'obiettivo principale. Nel contesto di questo progetto, l'obiettivo è acquisire i dati relativi ai gol effettivamente segnati e ai gol "teorici" derivati dall'indice xG di tutte le partite disputate durante la competizione UEFA EURO 2020. Questo primo step richiede la creazione di una tabella contenente *match_id*, risultato reale e risultato calcolato tramite l'indice xG.

La scelta di concentrarsi sulla competizione UEFA EURO 2020 è motivata dalla sua rilevanza nel panorama calcistico di alto livello, garantendo dati accurati e confrontabili con altri dataset. Inoltre, questa competizione ha offerto spunti interessanti per l'analisi del confronto tra risultati reali e quelli previsti dall'indice xG.

Il processo inizia con il caricamento di tutte le competizioni attraverso la funzione *sb.competitions()*, come precedentemente descritto.

Successivamente, per accedere alle partite della competizione desiderata, è necessario chiamare la funzione *sb.matches()*, fornendo gli ID corrispondenti alla competizione e alla stagione e salvando l'output in una variabile denominata "*matches*". In questo caso diventa *matches = sb.matches(competition_id=55, season_id=43)*.

64	12	86	Italy	Serie A	male
65	55	43	Europe	UEFA Euro	male
66	35	75	Europe	UEFA Europa League	male

Figura 13: Parametri necessari per accedere ai dati di UEFA EURO 2020

Ora, all'interno della variabile "*matches*", sono presenti tutte le partite del campionato europeo. Tuttavia, per il nostro scopo, sono necessarie solo alcune informazioni *(match_id* e risultato).
Inoltre, è necessario creare colonne per gli xG di ciascuna delle due squadre. Questo è un dato che StatsBomb non fornisce ma che è possibile ricavarsi. Infatti, l'Expetced Goals totale di una squadra al termine di una partita non è altro che la somma del valore xG di ogni singolo tiro effettuato da quella squadra. Per esempio: se una squadra effetua 3 tiri con indice xG pari a 0,33 l'indice xG totale della squadra a fine partita è 1, di conseguenza la squadra "teoricamente" dovrebbe segnare 1 gol ipotizzando che 1 dei 3 tiri si trasformi in xgol.

Per calcolare l'xG totale, in termini di linguaggio Python, è necessario fare un ciclo for iterando sulle righe del dataframe salvato nella variabile *matches* e, dopo aver creato le colonne *xG_home* e *xG_away*, sommare i valori xG dei tiri di ciascuna delle due squadre. Dopo averli sommati il valore totale finale viene aggiunto alla relativa colonna.

```python
# Creare un elenco per contenere i risultati
results = []

# Iterare sulle righe del DataFrame matches
for index, match in matches.iterrows():
    match_id = match['match_id']
    home_team = match['home_team']
    away_team = match['away_team']
    home_score = match['home_score']
    away_score = match['away_score']

    # Convertire match_id in un intero
    match_id_int = int(match_id)

    # Calcolare l'Expected Goals (xG) per le squadre di casa e ospiti
    events = sb.events(match_id_int)
    xG_home = events[events['team'] == home_team]['shot_statsbomb_xg'].sum()
    xG_away = events[events['team'] == away_team]['shot_statsbomb_xg'].sum()

    # Aggiungere i risultati alla lista
    results.append({
        'match_id': match_id_int,
        'home_score': home_score,
        'away_score': away_score,
        'xg_home': xG_home,
        'xg_away': xG_away
    })

# Creare un DataFrame con i risultati
results_df = pd.DataFrame(results)

# Visualizzare il DataFrame
results_df
```

Figura 14: Codice Python per ottenere il dataframe con gli xG totali

Il codice Python illustrato nella Figura 14 mostra come calcolare gli xG totali per ciascuna squadra in ogni partita. Successivamente, sono state aggiunte colonne aggiuntive ("*delta_home*" e "*delta_away*") per visualizzare la differenza tra i gol effettivi e il valore xG, anche se tali colonne non sono strettamente necessarie per il progetto.

```
# Aggiungi le colonne delta_home e delta_away
results_df['delta_home'] = results_df.apply(lambda row: row['home_score'] - row['xg_home'], axis=1)
results_df['delta_away'] = results_df.apply(lambda row: row['away_score'] - row['xg_away'], axis=1)

# Visualizza il DataFrame aggiornato
results_df
```

	match_id	home_score	away_score	xg_home	xg_away	delta_home	delta_away
0	3795108	1	1	3.957757	6.928069	-2.957757	-5.928069
1	3788769	1	4	1.090950	2.330365	-0.090950	1.669635
2	3788766	1	0	1.669124	0.240743	-0.669124	-0.240743
3	3795220	1	1	4.315651	4.536130	-3.315651	-3.536130
4	3788761	1	0	1.583975	0.553248	-0.583975	-0.553248
5	3788764	2	4	1.935741	2.130545	0.064259	1.869455

Figura 15: dataframe finale

Il risultato finale è un dataset più intuitivo e concentrato, pronto per essere utilizzato nell'analisi statistica successiva.

2.3: Assegnazione dei punti a ciascuna squadra

Dopo aver ottenuto un dataset soddisfacente, l'attenzione si sposta sull'assegnazione dei punti a ciascuna squadra, un passaggio essenziale per la successiva analisi statistica e la creazione di una classifica basata sui risultati reali e su quelli stimati tramite l'indice xG.

Inizialmente, i dati sono stati trasferiti su un foglio di calcolo Excel, dove ogni match_id è stato associato alle due squadre coinvolte, ai risultati in termini di Expected Goals e ai risultati reali. Successivamente, le squadre che hanno vinto sono state evidenziate in verde, mentre quelle che hanno pareggiato sono state contrassegnate in giallo.

GIRONE F							
UNGHERIA	PORTOGALLO	3788752	0	3		0,197621	2,681034
FRANCIA	GERMANIA	3788751	1	0		0,446024	0,825419
UNGHERIA	FRANCIA	3788763	1	1		0,473493	1,690028
PORTOGALLO	GERMANIA	3788764	2	4		1,935741	2,130545
PORTOGALLO	FRANCIA	3788773	2	2		2,042418	1,506618
GERMANIA	UNGHERIA	3788774	2	2		2,097695	0,737135
OTTAVI DI FINALE							
GALLES	DANIMARCA	3794689	0	4		0,598431	2,155293
ITALIA	AUSTRIA	3794685	2	1	dts	1,517069	1,038546
PAESI BASSI	REP. CECA	3794690	0	2		0,838034	1,273121
BELGIO	PORTOGALLO	3794687	1	0		0,224336	1,214752
CROAZIA	SPAGNA	3794686	3	5	dts	2,788224	2,780003
FRANCIA	SVIZZERA	3794691	3 (7)	3 (8)	dcr	6,56339	5,771544
INGHILTERRA	GERMANIA	3794688	2	0		1,163511	1,058548
SVEZIA	UCRAINA	3794692	1	2	dts	0,911922	1,344288

Figura 16: Foglio Excel per l'assegnazione dei punteggi

Ho optato per l'assegnazione di punteggi differenziati a seconda della fase della competizione:

Fase a gironi:

- Vittoria: 3 punti
- Pareggio: 1 punto
- Sconfitta: 0 punti

Fase ad eliminazione diretta:

- Vittoria nei 90 o 120 minuti: 3 punti
- Sconfitta nei 90 o 120 minuti: 0 punti
- Vittoria ai calci di rigore: 2 punti
- Sconfitta ai calci di rigore: 1 punto

Per quanto riguarda i punti assegnati basandosi sul valore xG, sono stati attribuiti 3 punti alla squadra con il valore maggiore e 0 punti a quella con il valore inferiore.

L'analisi dei risultati ha prodotto due classifiche: una basata sui risultati reali e l'altra basata sui punti assegnati analizzando i valori xG delle partite.

CLASSIFICHE REALI			CLASSIFICHE XG	
ITALIA	19		ITALIA	15
GALLES	4		GALLES	3
SVIZZERA	7		SVIZZERA	6
TURCHIA	0		TURCHIA	0
BELGIO	12		BELGIO	12
DANIMARCA	9		DANIMARCA	15
FINLANDIA	3		FINLANDIA	0
RUSSIA	3		RUSSIA	3
PAESI BASSI	9		PAESI BASSI	9
AUSTRIA	6		AUSTRIA	6
UCRAINA	6		UCRAINA	6
MACEDONIA	0		MACEDONIA	0
INGHILTERRA	17		INGHILTERRA	18
CROAZIA	4		CROAZIA	3
REP. CECA	7		REP. CECA	6
SCOZIA	1		SCOZIA	6
SVEZIA	7		SVEZIA	3
SPAGNA	11		SPAGNA	15
SLOVACCHIA	3		SLOVACCHIA	0
POLONIA	1		POLONIA	6
FRANCIA	6		FRANCIA	9
GERMANIA	4		GERMANIA	6
PORTOGALLO	4		PORTOGALLO	6
UNGHERIA	2		UNGHERIA	0

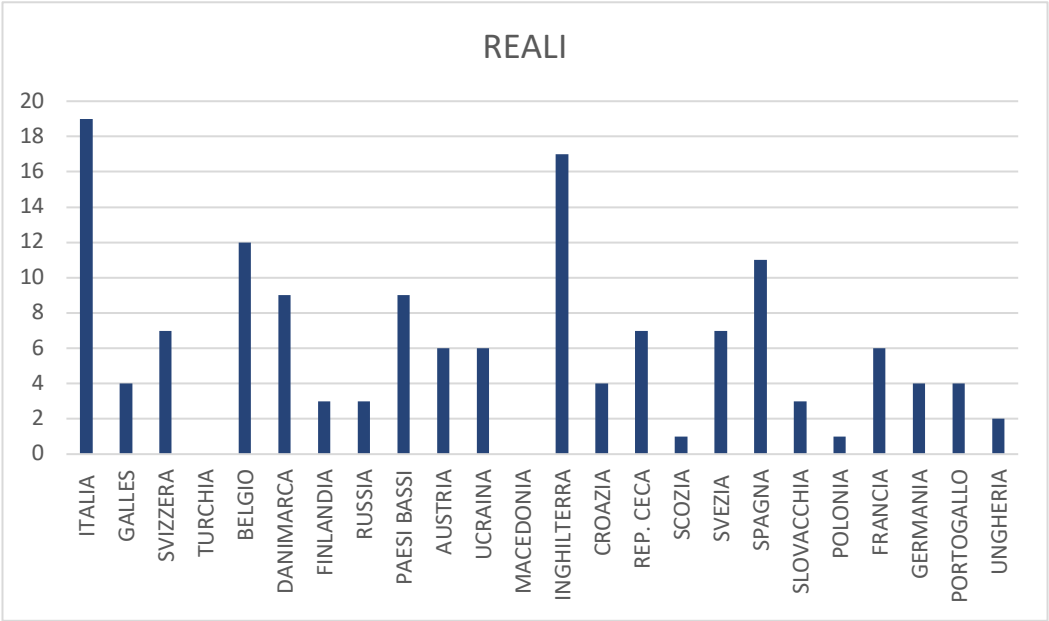
Figura 17: Classifica punti reali e punti xG

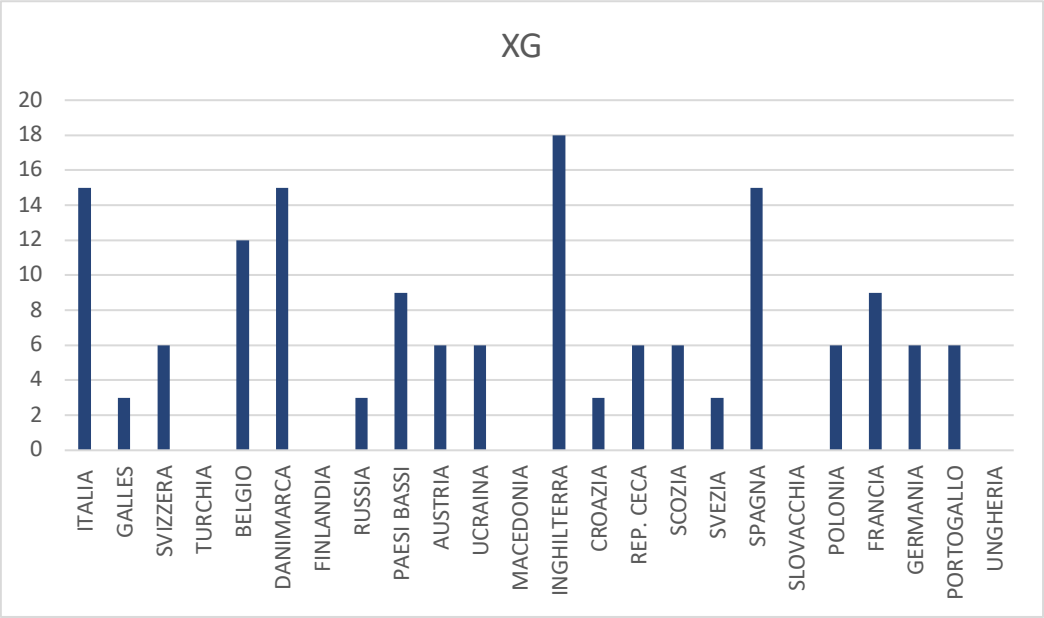
Un'occhiata ai punteggi "reali" rivela l'Italia e l'Inghilterra come le squadre con il maggior numero di punti, in linea con il loro cammino fino alla finale. Tuttavia, la classifica xG mostra emergere altre tre squadre in testa: Danimarca, Belgio e Spagna.

La Danimarca dati alla mano è la squadra che ha raccolto meno in confronto a quanto prodotto. Nella fase a gironi, infatti, in tutte e 3 le partite ha avuto l'indice xG superiore alla squadra avversaria ma, anziché totalizzare 9 punti, ne ha totalizzati solamente 3.

L'esempio opposto è rappresentato dall'andamento nel girone della formazione svedese. La Svezia ha raccolto 7 punti nel girone sebbene abbia prevalso in termini di xG solamente nella partita contro la Slovacchia. Questo indica che è una squadra molto precisa in zona tiro.

Questi esempi pratici evidenziano il quesito centrale della tesi: "Qual è il rapporto tra i gol effettivamente realizzati e il valore xG in una partita/competizione? C'è una differenza statisticamente significativa?"





Tale interrogativo, aiutandoci con questi due istogrammi e con altri strumenti statistici, sarà approfondito nei prossimi capitoli.

Capitolo 3

Applicazione dei test statistici e analisi dei dati

Questo capitolo si focalizza sulla fase chiave del progetto, che è l'applicazione di test statistici per rispondere al quesito di tesi. Nella selezione di un metodo appropriato, ho optato per il test di Kolmogorov-Smirnov. Questa scelta è stata guidata dalla necessità di valutare la differenza tra le distribuzioni reali e xG in modo non parametrico, senza fare specifiche assunzioni sulla forma delle distribuzioni. C'erano altri test che potevano essere considerati, tra cui il t-test e il z-test, ma il test di Kolmogorov-Smirnov è emerso come la scelta più adatta per il nostro contesto di analisi.

3.1: Introduzione al test di Kolmogorov-Smirnov

Il test di Kolmogorov-Smirnov è un metodo statistico utilizzato per confrontare un campione di dati con una distribuzione teorica o per confrontare due campioni di dati[18]. A differenza di altri test, non richiede la suddivisione preventiva dei dati in classi di frequenza e si basa sulla frequenza cumulativa relativa dei dati. Questo test valuta la differenza massima tra le distribuzioni cumulative empiriche dei due campioni. È un test non parametrico e non richiede assunzioni specifiche sulla forma delle distribuzioni[19]. Il test può essere applicato anche a campioni di dimensioni ridotte. Tuttavia, è sensibile alle differenze nella zona centrale dei dati, ma meno efficace nel rilevare differenze nelle code delle distribuzioni.

[18] https://www.wikiwand.com/it/Test_di_Kolmogorov-Smirnov
[19] https://it.wikipedia.org/wiki/Test_di_Kolmogorov-Smirnov

Nel nostro caso, l'ipotesi nulla afferma che non esiste una differenza statistica tra le due distribuzioni, mentre l'ipotesi alternativa sostiene il contrario. Per respingere l'ipotesi nulla, è necessario ottenere un p-value inferiore a 0,05, corrispondente a un livello di significatività del 5%.

3.2: Test di KS applicato sulle distribuzioni per punti: gruppo unico

I test possono essere applicati alle distribuzioni basate sui punti assegnati (classifiche calcolate in precedenza) o al confronto tra gol reali e gol stimati tramite xG.

Il primo test condotto è stato volto a verificare se esistesse una differenza statisticamente significativa tra le distribuzioni dei punti ottenuti. In altre parole, l'obiettivo era valutare se la differenza tra le due classifiche fosse significativa.

Quindi:

Media punti reali: 6,04166667
Media punti xG: 6,375

Ipotesi Nulla (H0): La distribuzione dei punti ottenuti realmente *non è uguale* alla distribuzione dei punti ipoteticamente ottenuti con gli xG

Ipotesi Alternativa (H1): La distribuzione dei punti ottenuti realmente *è uguale* alla distribuzione dei punti ipoteticamente ottenuti con gli xG

Applicando il test di Kolmogorov-Smirnov, otteniamo:

Test di KS:
Statistiche del test di Kolmogorov-Smirnov: 0.125
P-value: 0.994161229482218

Si osserva che il P-value è nettamente superiore a 0,05, quindi non siamo in grado di rifiutare l'ipotesi nulla (H0).

3.3: Test di KS applicato sulle distribuzioni per punti: 3 gruppi

Per aumentare l'affidabilità del test abbiamo anche applicato il test basato sulle distribuzioni per punti suddividendo le squadre in 3 gruppi. Le 24 squadre son state suddivise in 3 gruppi "meritocratici": nel primo gruppo (gruppo A) son state inserite le 8 squadre qualificatesi ai quarti di finale, nel secondo gruppo (gruppo B) le 8 squadre le squadre eliminate agli ottavi e nel terzo gruppo (gruppo C) le 8 nazioni eliminate nella fase a gironi.

Ottenendo così i seguenti gruppi:
Gruppo A: Svizzera, Belgio, Rep. Ceca, Ucraina, Spagna, Italia, Danimarca, Inghilterra
Gruppo B: Galles, Austria, Paesi Bassi, Portogallo, Croazia, Francia, Germania, Svezia
Gruppo C: Turchia, Finlandia, Russia, Macedonia, Scozia, Slovacchia, Polonia, Ungheria

A questo punto abbiamo applicato il test di Kolmogorov-Smirnov confrontando i punti reali e i punti xG per ogni singolo gruppo.

Test gruppo A:

Media punti reali: 11
Media punti xG: 11,625

Ipotesi Nulla (H0): La distribuzione dei punti ottenuti realmente dal gruppo A *non è uguale* alla distribuzione dei punti ipoteticamente ottenuti con gli xG dal gruppo A

Ipotesi Alternativa (H1): La distribuzione dei punti ottenuti realmente dal gruppo A *è uguale* alla distribuzione dei punti ipoteticamente ottenuti con gli xG dal gruppo A

Applicando il test otteniamo:

Test di KS:
Statistiche del test di Kolmogorov-Smirnov: 0.25
P-value: 0.9639

Test gruppo B:

Media punti reali: 5,5
Media punti xG: 5,625

Ipotesi Nulla (H0): La distribuzione dei punti ottenuti realmente dal gruppo B *non è uguale* alla distribuzione dei punti ipoteticamente ottenuti con gli xG dal gruppo B

Ipotesi Alternativa (H1): La distribuzione dei punti ottenuti realmente dal gruppo B *è uguale* alla distribuzione dei punti ipoteticamente ottenuti con gli xG dal gruppo B

Applicando il test otteniamo:

Test di KS:
Statistiche del test di Kolmogorov-Smirnov: 0.375
P-value: 0.6272

Test gruppo C:

Media punti reali: 1,625

Media punti xG: 1,875

Ipotesi Nulla (H0): La distribuzione dei punti ottenuti realmente dal gruppo C *non è uguale* alla distribuzione dei punti ipoteticamente ottenuti con gli xG dal gruppo C

Ipotesi Alternativa (H1): La distribuzione dei punti ottenuti realmente dal gruppo C *è uguale* alla distribuzione dei punti ipoteticamente ottenuti con gli xG dal gruppo C

Applicando il test otteniamo:

Test di KS:
Statistiche del test di Kolmogorov-Smirnov: 0.375
P-value: 0.6272

Si può osservare che, in nessuno dei 3 gruppi, è stato possibile rifiutare l'ipotesi nulla H0.

3.4: Test di KS applicato sulle distribuzioni per goal: gruppo unico

Dopo aver applicato il test di Kolmogorov-Smirnov alle distribuzioni basate sui punti, per ottenere ulteriori conferme rispetto ai test precedenti, abbiamo deciso di eseguire gli stessi test, ma su distribuzioni basate sui gol reali e sugli Expected Goals.

Avremo, quindi, una distribuzione composta dal totale dei gol effettivamente realizzati dalle squadre e un'altra composta dal valore xG accumulato durante l'arco del girone.
Il primo test è stato effettuato, come in precedenza, sul gruppo unico. Quindi:

Media goal reali: 3,916666667
Media xG: 3,881763333

Ipotesi Nulla (H0): La distribuzione dei goal realmente realizzati *non è uguale* alla distribuzione dei gol "teoricamente" realizzati basati sul valore xG

Ipotesi Alternativa (H1): La distribuzione dei goal realmente realizzati *è uguale* alla distribuzione dei gol "teoricamente" realizzati basati sul valore xG

Applicando il test di Kolmogorov-Smirnov, otteniamo:

Test di KS:
Statistiche del test di Kolmogorov-Smirnov: 0.25
P-value: 0.4490368220409109

Si osserva che il P-value è nettamente superiore a 0,05, quindi non siamo in grado di rifiutare l'ipotesi nulla (H0).

3.5: Test di KS applicato sulle distribuzioni per goal: 3 gruppi

Infine, l'ultima verifica da effettuare riguarda le distribuzioni basate sui gol, ma con le squadre suddivise in 3 gruppi, come descritto nel test del paragrafo 3.3.

I 3 gruppi sono gli stessi del paragrafo 3.3 e i risultati sono i seguenti:

Test gruppo A:

Media punti reali: 4,75
Media punti xG: 4,45811125

Ipotesi Nulla (H0): La distribuzione dei goal effettuati realmente dal gruppo A *non è uguale* alla distribuzione dei gol "teoricamente" realizzati basati sul valore xG del gruppo A

Ipotesi Alternativa (H1): La distribuzione dei goal effettuati realmente dal gruppo A *è uguale* alla distribuzione dei gol "teoricamente" realizzati basati sul valore xG del gruppo A

Applicando il test otteniamo:

Test di KS:
Statistiche del test di Kolmogorov-Smirnov: 0.25
P-value: 0.98010878010878

Test gruppo B:

Media punti reali: 5

Media punti xG: 4,6914595

Ipotesi Nulla (H0): La distribuzione dei goal effettuati realmente dal gruppo B *non è uguale* alla distribuzione dei gol "teoricamente" realizzati basati sul valore xG del gruppo B

Ipotesi Alternativa (H1): La distribuzione dei goal effettuati realmente dal gruppo B *è uguale* alla distribuzione dei gol "teoricamente" realizzati basati sul valore xG del gruppo B

Applicando il test otteniamo:

Test di KS:
Statistiche del test di Kolmogorov-Smirnov: 0.375
P-value: 0.6601398601398599

Test gruppo C:

Media punti reali: 2
Media punti xG: 2,49571925

Ipotesi Nulla (H0): La distribuzione dei goal effettuati realmente dal gruppo C *non è uguale* alla distribuzione dei gol "teoricamente" realizzati basati sul valore xG del gruppo C

Ipotesi Alternativa (H1): La distribuzione dei goal effettuati realmente dal gruppo C *è uguale* alla distribuzione dei gol "teoricamente" realizzati basati sul valore xG del gruppo C

Applicando il test otteniamo:

Test di KS:
Statistiche del test di Kolmogorov-Smirnov: 0.375
P-value: 0.6601398601398599

Anche in questo caso, come evidenziato nello stesso test con la distribuzione per punti, non è stato possibile rifiutare l'ipotesi nulla (H0) poiché i P-value sono maggiori del livello di significatività.

Capitolo 4

Interpretazione dei risultati e conclusioni

L'analisi dettagliata condotta attraverso una serie di esperimenti e test statistici mirava a valutare se ci fossero differenze significative tra le distribuzioni dei punti, dei gol reali e degli Expected Goals (xG) delle squadre nel contesto della competizione UEFA EURO 2020. Abbiamo adottato diversi approcci metodologici, tra cui l'applicazione del test di Kolmogorov-Smirnov per esaminare le distribuzioni sotto vari aspetti.

I risultati ottenuti non hanno mai fornito evidenze statisticamente significative per rifiutare l'ipotesi nulla, che sostiene l'assenza di differenze sostanziali tra le distribuzioni analizzate.

L'ipotesi nulla, che afferma l'uguaglianza tra i fenomeni misurati, è stata costantemente supportata dai p-value ottenuti nei test di Kolmogorov-Smirnov. Questi valori, superando il livello di significatività del 5%, indicano che non vi è sufficiente evidenza statistica per concludere che le distribuzioni dei punti, dei gol reali e degli Expected Goals siano diverse.

In sintesi, l'insieme dei miei esperimenti e di tutti i test fatti non ha mai messo in evidenza in modo significativo alcuna disparità sostanziale tra le distribuzioni dei risultati effettivi e quelli previsti dagli Expected Goals. Questi risultati contribuiscono alla comprensione più ampia del rapporto tra i risultati reali e quelli previsti attraverso l'analisi avanzata nel contesto del calcio e della UEFA EURO 2020.

Elenco figure

Bibliografia e Sitografia

[1]: V. Gouveia et al., Notational Analysis on Goal Scoring and Comparison in Two of the Most Important Soccer Leagues: Spanish La Liga and English Premier League, Appl. Sci. 2023, 13(12), 6903; https://doi.org/10.3390/app13126903

[2]: Kubay et al., Trends of Goal Scoring Patterns in Soccer: A Retrospective Analysis of Five Successive FIFA World Cup Tournaments, J Hum Kinet. 2019 Oct; 69: 231–238. doi: 10.2478/hukin-2019-0015

[3]: B. Micovic et al., The Qatar 2022 World Cup warm-up: Football goal-scoring evolution in the last 14 FIFA World Cups (1966–2018), Front. Psychol., 2023, Volume https://doi.org/10.3389/fpsyg.2022.954876

[1] https://www.ilsole24ore.com/art/data-analytics-sport-mercato-3-miliardi-dollari-AEmxJepB

[2] https://www.rainews.it/articoli/2023/05/pallacanestro-basket-olimpia-ettore-messina-gregg-popovich-san-antonio-spurs-big-data-statistica-cdcdb120-1d3e-4c83-8557-5ecb49aa953a.html

[3] https://www.hitrech.com/spogliatoio/il-gps-nel-calcio-e-davvero-la-migliore-tecnologia.html

[4] https://medium.com/calcio-datato/quali-sono-le-zone-di-campo-pi%C3%B9-battute-dalle-squadre-di-serie-a-3c88c76625b7

[5] https://it.whoscored.com/Matches/1746244/

[6] https://www.wired.it/scienza/lab/2016/03/25/calcio-big-data/

[7] https://wylab.net/formazione/football-data-analyst/

[8] https://www.mancity.com/club/partners/acronis

[9] https://strategiedigitali.net/kpi-aziendali-definizione-significato-ed-esempi/

[10] https://www.acronis.com/en-eu/sports/manchester-city/

[11] https://www.ultimouomo.com/expected-goals/

[12] https://www.90min.com/it/posts/cosa-sono-e-come-calcolare-gli-expected-goal-xg

[13] https://www.eurosport.it/calcio/cosa-sono-gli-expected-goals-e-a-cosa-servono-le-5-statistiche-avanzate-piu-utilizzate-per-capire-me_sto9204614/story.shtml

[14] https://thegegenpress.it/glossario-statistico/

[15] https://www.statsperform.com/resource/expected-assists-in-context-2/

[16] https://statsbomb.com/what-we-do/soccer-data/360-2/

[17] https://www.ibm.com/it-it/topics/api

[18] https://www.wikiwand.com/it/Test_di_Kolmogorov-Smirnov

[19] https://it.wikipedia.org/wiki/Test_di_Kolmogorov-Smirnov

Ringraziamenti

Innanzitutto, un ringraziamento particolare va al mio relatore, il Professor Marco Roccetti, persona dotata di umanità e competenza tecnica straordinarie. Grazie per aver accettato come tesista, per avermi costantemente motivato rendendomi consapevole delle mie potenzialità, per avermi dedicato tempo e pazienza, e per aver reso stimolante l'argomento che ho affrontato.

Desidero ringraziare i miei genitori che, grazie ai loro insegnamenti, mi hanno reso la persona che sono oggi. Grazie per avermi spronato sempre a fare qualcosa in più e grazie per avermi regalato una splendida sorella.
Grazie, quindi, a mia sorella per il costante affetto e supporto su cui posso sempre contare nonostante la distanza e il poco tempo che trascorriamo insieme.

Vorrei ringraziare tutti i nonni, fisicamente presenti e non, che non mi hanno mai fatto mai mancare nulla. Un grazie particolare al nonno Mario per tenermi aggiornato sui pericoli dei posti in cui vado. Grazie anche a tutto il resto della famiglia: zii, cugini e tutti quanti. È davvero una fortuna poter contare su una famiglia così. Grazie Bika per le ripetizioni di economia aziendale: ho preso 28 ma conto economico e stato patrimoniale non li ho mai capiti.

Un grande ringraziamento va anche ai miei amici, da quelli che vedo e sento quotidianamente fino a tutti quelli che hanno condiviso solo una piccola parte di questo percorso con me. Grazie per avermi supportato e sopportato, so bene di avere un carattere difficile. Grazie Jack per avermi scarrozzato in Punto nera a 5000 giri per anni, grazie Rubo per appoggiarmi quando voglio andare a casa presto, grazie Faggio per non avermi mai appeso ad una grata di un campo da

padel, grazie Gio per avermi inviato con la tua solita velocità il frontespizio. Non vi ho citati tutti personalmente solo per una questione di spazio.
Grazie anche ai miei amici delle superiori, Salta e Eugert in particolare: siete amici veri su cui so di poter contare sempre.

Dedico a tutti quanti questa tesi, mi auguro che il vostro prezioso supporto non venga mai a mancare.