2023/24

CAPOLAVORO: Football Data Science

Gianmarco Roberti

Capolavoro

2023/24

Indice

Introduzione

Negli ultimi tempi mi sono molto appassionato alla data analysis e quindi ho deciso di creare il mio capolavoro proprio su questo argomento, intrecciandolo con una delle mie più grandi passioni, il calcio. L'obiettivo di questo progetto era quello di apprendere nuove nozioni di Excel, SQL, Python, ma anche proprio mettere mano ad un vero dataset preso da Kaggle e altri siti calcistici come https://www.fantacalcio.it/statistiche-serie-a/2023-24/statistico e https://fbref.com/it/comp/11/stats/Statistiche-di-Serie-A e usare nuovi ambienti di sviluppo ovvero Google Colab e Looker Studio (entrambi di proprietà di Google).

Idea

L’idea per questo progetto mi è venuta guardando alcuni video sul web di alcuni Data Analyst professionisti e io ho cercato di riportare il loro lavoro nel mio mondo. Infatti la mia idea è scaturita dove aver visto alcuni video su YouTube di @JM Dieke. La mia idea comprendeva inizialmente la creazione di un file .xlsx dove andavano riportati tutti i dati presi da Internet, successivamente andava realizzata la dashboard per la visualizzazione di questi dati, facendo anche le opportune analisi. Le analisi comprendono principalmente il numero di gol fatti, quello degli assist effettuati, dei cartellini ricevuti, dei gol attesi e delle nazioni di provenienza dei calciatori. I dati da me trattati comprendono le statistiche della stagione 2023/24, quindi i dati non sono completi, poiché nel momento in cui sto scrivendo, mancano ancora 2 partite del campionato di Serie A.

Cosa ho studiato?

Per poter redigere questo progetto ho inizialmente cercato del materiale online per quanto riguarda la data analysis, in particolare la football analysis. Ho trovato alcuni progetti su GitHub su questo argomento e ho poi approfondito il codice Python utile all'analisi dei dati tramite dei video su Numpy, Pandas e Matplotlib su YouTube. Per seguire i corsi su queste librerie di Python, mi sono affidato a delle playlist sul canale YouTube di @EdoardoMidali, in cui ho seguito non solo tutorial su questo argomento ma anche altri video su altri linguaggi di programmazione.

Successivamente mi sono informato anche per quanto riguarda Google Colab ed ho scoperto che è una piattaforma basata su cloud che consente di scrivere e eseguire codice Python attraverso il tuo browser. Ultimamente molti utenti stanno usando questi strumenti poiché sono gratuiti e offrono GPU e RAM a tutti gli utenti che hanno un account Google.

Infine ho iniziato a prendere confidenza con Looker Studio, uno strumento gratuito, offerto da Google, che trasforma i dati in dashboard e report informativi, facili da leggere e condividere e completamente personalizzabili. Ho imparato ad usare questo servizio sempre grazie sia alla visione di video su YouTube ma anche grazie all'enorme mole di informazioni che girano in rete. Grazie a questo programma sono riuscito a realizzare una dashboard intuitiva ma efficace.

Ovviamente, per questo progetto, mi servivano anche delle nozioni di Excel, di cui però avevo una buona base avendolo studiato durante il mio percorso scolastico ma anche personale, facendo anche ore di ripetizione a ragazzi non molto ferrati sull'argomento.

Realizzazione progetto

Questo progetto è stato ideato e realizzato come "Capolavoro" per l'anno scolastico 2023/2024. Ho iniziato la realizzazione del progetto nel mese di Gennaio 2024, dopo aver seguito alcuni corsi su argomenti come la Data Analysis. Possiamo dividere il progetto in due parti, la prima che riguarda Excel, mentre l'altra riguarda Looker Studio:

1. Il mio dataset di partenza, ovvero l'insieme dei dati su cui poi ho effettuato alcune analisi, è stato scritto in un file .xlsx e comprendeva alcuni campi come l'id del calciatore, il cognome del calciatore, la squadra di appartenenza, la nazionalità del calciatore, il numero di gol effettuati, il numero di assist effettuati, i minuti giocati, le partite giocate, la somma dei gol e degli assist, gli expected goal, il numero di ammonizioni, il numero di espulsioni e il ruolo del calciatore. I dati, come già anticipato, si riferiscono alla stagione 2023/24. Nel file Excel ho quindi riportato tutti i dati che ho trovato sui siti prima citati. I dati trovati erano "grezzi", quindi prima di fare qualsiasi analisi, ho effettuato alcune operazioni di data cleaning, ovvero ho eliminato i dati che non mi servivano, quelli che erano errati e quelli che non erano formattati correttamente. Ad esempio i dati sulle nazioni avevano sigle non conformi e quindi li ho modificati e riadattati al mio progetto. Inoltre erano presenti nomi doppi, quindi, tramite alcune funzioni di Excel, ho eliminato i dati ripetuti.

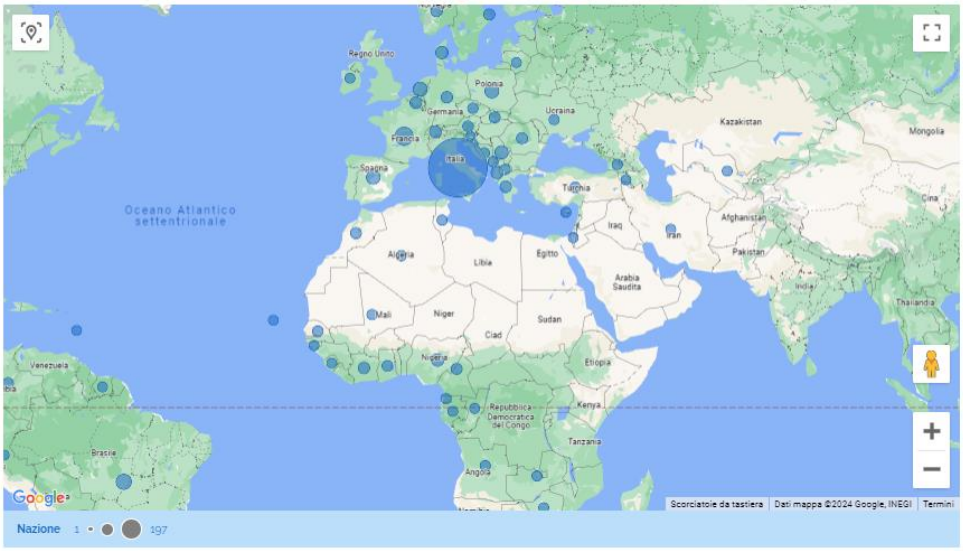
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	**Pos.**	**Giocatore**	**Nazione**	**Ruolo**	**Squadra**	**Età**	**Nato**	**PG**	**Tit**	**Min**	**Reti**	**Assist**	**G+A**	**Amm.**	**Esp.**	**xG**	**npxG**	**xAG**
2	1	Tammy Abraham	ENG	Att	Roma	26	1997	6	1	172	1	0	1	1	0	0,5	0,5	0,1
3	2	Francesco Acerbi	ITA	Dif	Inter	36	1988	27	24	2.208	3	1	4	1	0	1,6	1,6	1,7
4	3	Yacine Adli	FRA	Cen	Milan	23	2000	23	17	1.375	1	2	3	3	0	0,3	0,3	2,1
5	4	Michel Aebischer	SUI	Cen	Bologna	27	1997	35	25	2.140	0	1	1	7	0	1,1	1,1	1,5
6	5	Lucien Agoume	FRA	Cen	Inter	22	2002	1	0	5	0	0	0	0	0	0	0	0
7	6	Marley Aké	FRA	Att	Udinese	23	2001	1	0	10	0	0	0	0	0	0	0	0
8	7	Ebenezer Akinsanmiro	NGA	Cen	Inter	19	2004	1	0	15	0	0	0	0	0	0	0	0
9	8	Jean-Daniel Akpa-Akpro	CIV	Cen	Monza	31	1992	19	8	705	0	0	0	5	0	0,5	0,5	0,1
10	9	David Akpan Ankeye	NGA	Att	Genoa	22	2002	5	0	56	0	0	0	0	0			
11	10	Luis Alberto	ESP	Cen	Lazio	31	1992	32	29	2.292	5	7	12	7	0	4,7	4,7	4,7
12	11	Carlos Alcaraz	ARG	Cen	Juventus	21	2002	8	2	245	0	1	1	0	0	0,3	0,3	0,1
13	12	Pontus Almqvist	SWE	Att	Lecce	24	1999	29	24	2.052	2	1	3	5	0	2,3	2,3	2,6
14	13	Lorenzo Amatucci	ITA	Cen	Fiorentina	20	2004	2	0	22	0	0	0	0	0	0	0	0
15	14	Bruno Amione	ARG	Dif	Hellas Verona	22	2002	10	7	579	0	0	0	2	0	0,3	0,3	0
16	15	Felipe Anderson	BRA	Att	Lazio	31	1993	36	33	2.713	5	6	11	3	0	4,3	4,3	6,3
17	16	Angeliño	ESP	Dif	Roma	27	1997	14	12	971	0	0	0	2	0	0,1	0,1	1,1
18	17	Houssem Aouar	ALG	Cen	Roma	25	1998	15	8	687	3	0	3	3	0	2	2	0,4
19	18	Marko Arnautović	AUT	Att	Inter	35	1989	26	4	746	3	3	6	0	0	4,3	4,3	2,6
20	19	Kristjan Asllani	ALB	Cen	Inter	22	2002	22	6	764	1	2	3	1	0	0,6	0,6	1,1
21	20	Emil Audero	ITA	Por	Inter	27	1997	3	3	270	0	0	0	0	0	0	0	0
22	21	Tommaso Augello	ITA	Dif	Cagliari	29	1994	30	26	2.243	1	1	2	5	0	0,5	0,5	1,9
23	22	Carlos Augusto	BRA	Dif	Inter	25	1999	35	13	1.608	0	3	3	1	0	1	1	2,2
24	23	Yann Aurel Bisseck	GER	Dif	Inter	23	2000	15	8	814	2	0	2	0	0	0,9	0,9	0,5
25	24	Sardar Azmoun	IRN	Att	Roma	29	1995	22	3	564	3	0	3	4	0	2,9	2,9	1,2
26	25	Paulo Azzi	BRA	Dif	Cagliari	29	1994	25	9	950	0	0	0	1	0	1,2	1,2	0,7
27	26	Oussama El Azzouzi	MAR	Cen	Bologna	22	2001	16	4	442	2	2	4	3	0	0,5	0,5	0,6
28	27	Milan Badelj	CRO	Cen	Genoa	35	1989	32	31	2.455	1	3	4	6	0	1,4	1,4	1,1

Questo è una parte del mio file Excel in cui è contenuto il mio dataset per il progetto, poiché i giocatori totali analizzati sono circa 600.

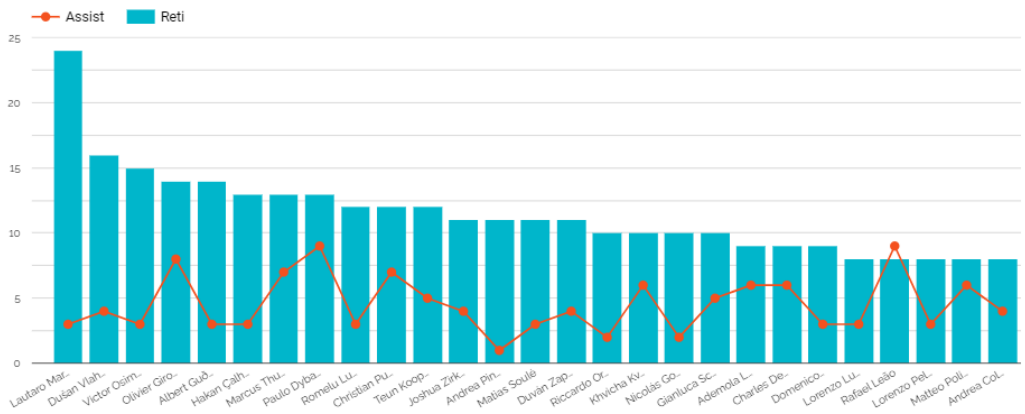
2. La seconda parte del progetto è stata svolta su Looker Studio, dove ho importato il file .csv con i dati del file Excel. Prima di effettuare questa operazione, utilizzando un convertitore online, ho convertito il file .xlsx in un file .csv di modo da poterlo caricare su Looker Studio. Successivamente ho caricato il file .csv nella sezione Dati di Looker Studio. Al mio report ho aggiunto un titolo "Serie A Players Dashboard" e, nella prima pagina, ho inserito una mappa per vedere la nazione di provenienza dei calciatori che giocano in Serie A. Da questo grafico è risultato che la maggior parte dei calciatori di Serie A provengono dall'Italia e generalmente dall'Europa. Fuori dai confini europei vediamo che ci sono molti calciatori argentini ma anche dell'Africa occidentale.
   Nella seconda pagina ho inserito un grafico combinato che conteneva i dati sui gol e sugli assist. In seguito ho inserito delle "schede" che mostrano alcuni dati sui calciatori, come la somma dei gol e degli assist, il massimo dei minuti totali giocati da un singolo calciatore, il numero massimo di ammonizioni prese da un calciatore e la somma di tutte le espulsioni.
   Nella terza pagina ho inserito un grafico a barre che mostra le reti effettivamente segnate e gli expected gol. Gli xG sono i gol previsti, ovvero una misura di quanti gol un calciatore avrebbe dovuto realizzare. Si può vedere come alcuni calciatori hanno over-performato, ovvero hanno segnato più gol di quelli attesi, mentre per altri erano attesi più gol.

Serie A Players Dashboard

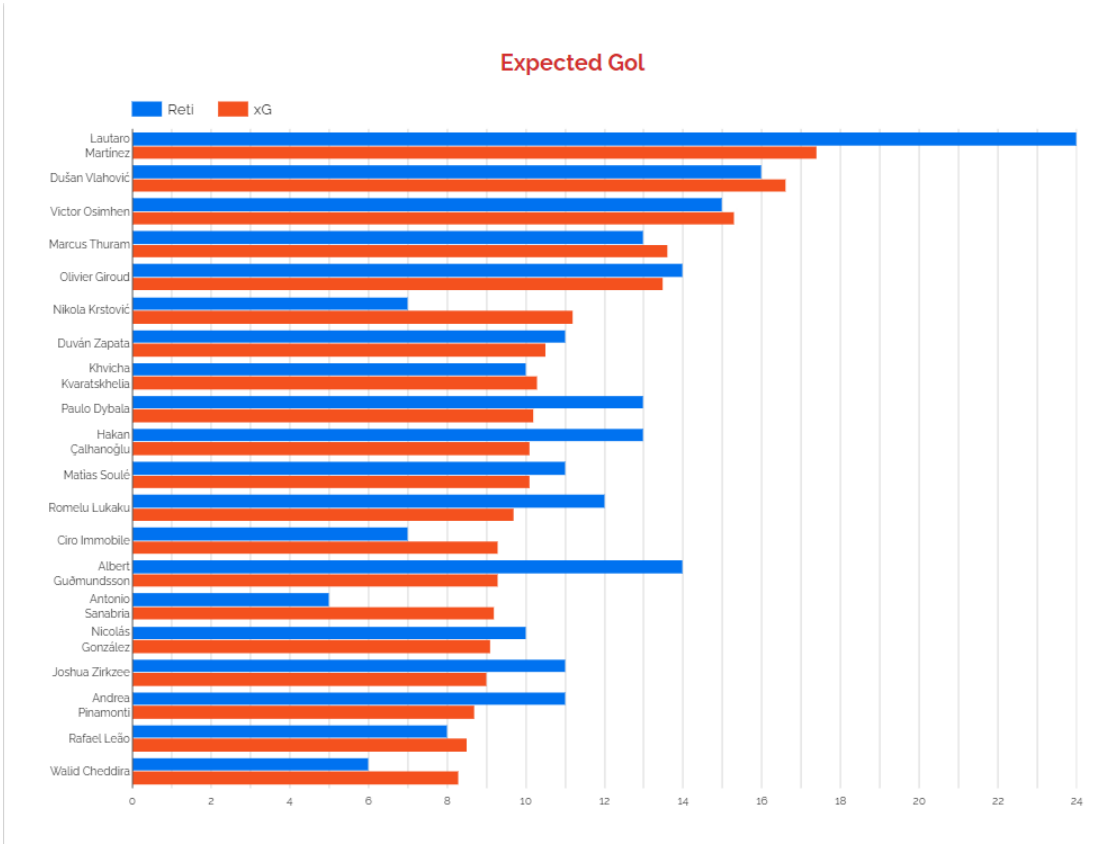
Da dove vengono i calciatori?



Gol e Assist



Gol+Assist	Minuti	Ammonizioni	Espulsioni
G+A	Min	Amm.	Esp.
27	3.240	14	60



Questa è la dashboard che contiene i dati sotto forma di grafici.

La dashboard inoltre è interattiva, questo significa che l'utente che andrà poi a consultare se ad esempio clicca, nel grafico misto, sulla barra relativa ai gol di un singolo calciatore, nei grafici sottostanti verranno visualizzate le statistiche relative a quel calciatore, quindi quanti gol ha segnato, qual è la sua media voto, quante ammonizioni a ricevuto e quale è il suo ruolo.
Invece se ad esempio l'utente, nel grafico a torta, clicca sulla porzione dedicata agli attaccanti visualizzerà solo i dati relativi agli attaccanti, quindi il numero massimo di gol realizzati da un attaccante, la media voto degli attaccanti, l'attaccante che ha ricevuto più cartellini gialli e il grafico misto riporterà le informazioni relative ai soli attaccanti.
Il link per accedere alla visualizzazione del report su Looker Studio è:
https://lookerstudio.google.com/s/oGVp5RJzlXI

Difficoltà

Le difficoltà riscontrate nella realizzazione di questo progetto sono state molteplici, innanzitutto ho lavorato su ambienti diversi da quelli che usavo abitualmente e quindi inizialmente mi sono sentito un po' spaesato. Successivamente ho avuto difficoltà nella gestione del Dataset e in alcune delle operazioni che dovevo effettuare su di esso, come, ad esempio, la creazione dei grafici per effettuare le analisi. Ad esempio ho avuto qualche difficoltà nella creazione del grafico a mappa, per la visualizzazione della nazionalità dei calciatori. In particolare ho avuto problemi con le sigle delle nazioni, poiché il programma non le riconosceva correttamente. Per superare queste difficoltà ho utilizzato materiale online che mi spiegava come poter effettuare le operazioni che desideravo, nello specifico ho utilizzato molti video-tutorial presenti su YouTube.

Conclusioni

Realizzare questo progetto mi è piaciuto molto poiché combinava sia la mia passione per il calcio, che il mio interesse verso la Data Science, che spero in futuro di poter approfondire. Credo che questo semplice progetto con la sua interfaccia possa essere il punto di partenza dei miei studi di statistica. Infatti mi piacerebbe molto poter lavorare ancora con i dati ed estrapolare da essi informazioni interessanti, soprattutto se parliamo dell'ambito calcistico. Ho scelto questa tipologia di progetto poiché penso che esso rappresenti il mio "ponte" dalla scuola superiore all'università: infatti durante la scuola secondaria di 2° grado ho studiato Informatica, e quindi anche Python, mentre all'università mi piacerebbe frequentare il corso di laurea in scienze statistiche, poiché è una delle materie che più mi affascina e che credo per il futuro sia tra le più fondamentali.