Data Analysis Report

Giuseppe Marino

24 giugno 2024

1 Introduzione

In questo report, espongo le scelte implementative e le strategie adottate nell' analisi dei dati, delineando l'efficacia e l'intuizione fornita dall'approccio utilizzato.

2 Soluzione: Metodologia e Approccio Concettuale

L' approccio all'analisi dei dati `e stato sistematico e guidato da un obiettivo ben definito: comprendere le strategie di gioco dei club calcistici esaminando i dataset disponibili per estrarre informazioni rilevanti che guidassero l'analisi verso la deduzione di risultati validi.

2.1 Analisi dei Club

L'obiettivo iniziale era identificare i club con le migliori performance e stili di gioco. Sono quindi partito dall'analisi dell'andamento storico dei dati, ponendo enfasi su quelli più recenti per riflettere le tendenze attuali senza trascurare l'evoluzione nel tempo.

2.2 Studio dei Giocatori:

Un'attenzione particolare è stata rivolta alla composizione delle squadre, valutando la distribuzione delle posizioni e l'età media dei giocatori per individuare tendenze nelle scelte di formazione e nella strategia di sviluppo del talento.

2.3 Valutazione dei Manager:

L'analisi si è estesa ai manager, confrontando le loro prestazioni per valutare l'impatto della loro leadership sulle prestazioni del club.

3 Soluzione: Descrizione Tecnica e Scelte Implementative

Ogni sezione di analisi è preceduta da un'introduzione che definisce gli obiettivi e conclude con un'interpretazione dei risultati ottenuti. Il Jupyter Notebook è strutturato in modo che ogni cella di codice sia accompagnata da commenti esplicativi per facilitare la comprensione del lettore.

3.1 Interazione con il Database e Scalabilità:

I dati sono stati importati e gestiti utilizzando Pandas. Anche se alcune informazioni erano incomplete, le analisi hanno fornito insight validi. L'approccio modulare adottato ha garantito che l'analisi rimanesse gestibile e scalabile, anche con dataset di grandi dimensioni. Per quanto riguarda la scalabilità, mentre l'utilizzo di Pandas e altre tecniche analitiche dimostra la capacità di gestire set di dati di ampia portata, la prova decisiva della loro efficacia si verifica nell'esecuzione in ambienti di produzione. La gestione della memoria e l'ottimizzazione delle prestazioni sono cruciali, specialmente quando si lavora con volumi di dati che raggiungono o superano il milione di record. Ad esempio, la funzione che introduce un ritardo nelle richieste di geocoding rappresenta una prudenza necessaria per evitare il superamento dei limiti imposti dai servizi di geolocalizzazione, ma potrebbe richiedere un adattamento per gestire carichi di lavoro di dimensioni maggiori. La correttezza dei risultati ottenuti è riconducibile alla loro conformità agli standard definiti in fase di preparazione dei dati.

3.2 Data Cleaning

Durante l'analisi, ho riscontrato occasionalmente inconsistenze nei dataset. Per affrontare questo problema, ho inizialmente agito in due metodi per poi produrre solo un'analisi seguendo il primo metodo. Esso prevedeva la selezione di dati coerenti e l'esclusione delle informazioni non corrispondenti tra i diversi set, garantendo così l'integrità dell'analisi senza pregiudicare l'efficienza del processo di elaborazione andando rimuovere le celle senza valori o a riempirle con dati generici, utilizzando i metodi "***df.dropna***" e "***df.fillna***" .
Il secondo metodo invece prevedeva la creazione di un jupyter notebook di nome "*dataCleaningExample*" nel quale ho pulito i dataset tramite l'utilizzo di una funzione "***def save_cleaned_data(dataframes, output_folder="cleanDatasets")*** " che per ogni dataframe utilizza il metodo "***df.fillna***. Fatto ciò, ho salvato i dataset puliti sempre attraverso la stessa funzione usata prima in una cartella denominata **cleanDatasets** nel percorso */IUMTWEB/DataAnalysis/cleanDatasets/.*
Durante l'utilizzo del secondo metodo ho potuto constatare che la pulizia dei datasets, fatta per intero e precedentemente all'analisi dati, aveva si prodotto gli effetti sperati, ovvero datasets completi e senza celle nulle, ma in un lasso di tempo raddoppiato rispetto all'utilizzo del primo metodo, che prevede invece la selezione di dati coerenti con il management diretto dei dataset dentro la stessa analisi, con un risparmio di tempo non indifferente. Per tali motivi, ho deciso di procedere con l'analisi dei dati utilizzando il primo metodo, lasciando però il secondo metodo visualizzabile per completezza.

3.3 Data Visualization:

I dati vengono visualizzati attraverso un'ampia gamma di sfumature grafiche, attentamente selezionate e variate per adattarsi a ogni contesto specifico. Sono stati utilizzati grafici evolutivi, radar e di geolocalizzazione, tra gli altri, per una rappresentazione ottimale. Ogni grafico è preceduto da una meticolosa preparazione e analisi dei dati, da cui vengono dedotti risultati pertinenti, fornendo contestualmente spiegazioni chiare e approfondite.

4 Documentazione del codice:

il codice è ben documentato, ogni passo che conduce all'analisi dei dati viene descritto e le varie celle sono tutte precedute da un markdown che descrive la funzionalità di quel codice.

5 Analisi dei Dati e Conclusioni:

Ho condotto un'analisi olistica, partendo da una visione di insieme per poi focalizzarmi su dettagli specifici dei club, dei giocatori e dei manager. Attraverso l'utilizzo di tecniche come il merge di Pandas, sono stato in grado di integrare e confrontare dati da prospettive diverse, il che ha migliorato la chiarezza dei risultati.
I risultati ottenuti forniscono intuizioni che possono essere applicate per migliorare le prestazioni dei club. Ad esempio, la focalizzazione su giocatori in specifici paesi o l'adozione di determinate formazioni di gioco può essere una strategia valida per i club in cerca di miglioramento.
In conclusione, questo report non solo evidenzia le prestazioni passate, ma offre anche una base su cui i club possono costruire per migliorare le loro future strategie di gioco e gestione.