UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI INGEGNERIA
Dipartimento di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria Informatica e dell'Automazione

TESI DI LAUREA

Progettazione e implementazione di una campagna di Data Analytics relativa ai risultati delle partite di calcio delle squadre nazionali

Design and implementation of a Data Analytics campaign regarding the match results of national soccer teams

Relatore

Prof. Domenico Ursino

Candidato

Aldo Paganica

Correlatore

Dott. Luca Virgili

ANNO ACCADEMICO 2023-2024

A mia mamma,
che è orgogliosa di me

Sommario

In un contesto sempre più digitalizzato e interconnesso, la Data Analytics assume un'importanza cruciale. I grandi volumi di dati disponibili oggi permettono di estrarre informazioni preziose che influenzano non solo le decisioni aziendali, ma anche le strategie future. Questo studio si propone di evidenziare l'utilità e l'efficacia della Data Analytics, dimostrando come l'analisi approfondita di grandi quantità di dati possa fornire una visione dettagliata dei fenomeni aziendali e offrire ulteriori linee guida per la pianificazione strategica. In questa tesi, abbiamo condotto una campagna di Data Analytics relativa ai risultati delle partite di calcio delle squadre nazionali. Dopo aver descritto i dati, abbiamo effettuato operazioni di ETL utilizzando il software Power BI per pulire, organizzare e predisporre i dati per le analisi successive. Attraverso l'uso di tecniche di Data Visualization, abbiamo creato report interattivi sui risultati delle partite e sui marcatori delle nazionali.

Keyword: Big Data, Data Analytics, Extract, Trasform and Load, Power BI, Data Visualization, Data Cleaning, Data Warehousing

Indice

Elenco delle figure

Introduzione

"If you can't measure it, you can't improve it."

(Peter Drucker)

Nell'era digitale attuale, l'importanza della misurazione e dell'analisi dei dati è diventata fondamentale per il successo aziendale. La citazione di Peter Drucker sottolinea come la capacità di misurare i dati sia essenziale per apportare miglioramenti e prendere decisioni informate. L'analisi dei Big Data si è imposta come uno strumento indispensabile nel mondo delle imprese, permettendo di attingere a un vasto panorama di informazioni per guidare strategie e operazioni.

La presente tesi esplora l'importanza della Big Data Analytics nel contesto delle moderne strategie aziendali, illustrando come l'abilità di interpretare e analizzare grandi quantità di dati possa portare a scoprire nuove opportunità e vantaggi competitivi.

L'era digitale ha visto una crescita esponenziale nella produzione di dati, noti come "Big Data". Gestire efficacemente questi dati richiede tecniche e metodologie avanzate, sviluppate per affrontare la complessità e la velocità con cui le informazioni vengono generate.

Le aziende moderne utilizzano i Big Data in molti settori delle loro operazioni, dal marketing alla finanza, dalla gestione delle risorse umane alla produzione. La capacità di raccogliere, analizzare e interpretare dati su larga scala è diventata un elemento fondamentale per il successo aziendale, permettendo di ottimizzare i processi decisionali e migliorare l'efficienza operativa.

L'emergere di nuove figure professionali, come i Data Analyst e i Data Scientist, riflette l'importanza della gestione e dell'analisi dei dati. Questi esperti sono responsabili della trasformazione dei dati in conoscenze pratiche, attraverso l'uso di tecniche avanzate di analisi e visualizzazione.

In conclusione, l'implementazione delle tecniche di analisi di Big Data è fondamentale in un mondo sempre più digitale e interconnesso. Il loro utilizzo strategico può rivelare nuove opportunità di business, potenziare i processi decisionali e perfezionare le strategie aziendali., contribuendo in modo significativo al successo e alla sostenibilità delle organizzazioni.

Alla luce di queste considerazioni, questo elaborato propone una campagna di Data Analytics relativa ai risultati delle partite di calcio delle squadre nazionali. Dopo un accurato processo di acquisizione dei dati, abbiamo eseguito operazioni di ETL (Extract, Transform, Load) per pulire, modellare e preparare i dati per le successive fasi di analisi. Tutti questi processi sono stati gestiti utilizzando il software di Business Intelligence Power BI.

In seguito, sono stati creati vari report interattivi relativi alle tabelle del nostro Dataset, che riguardano i risultati e i marcatori delle partite delle squadre nazionali.

La presente tesi è composta da cinque capitoli strutturati come di seguito specificato:

- Il Capitolo 1 fornisce una panoramica approfondita sulla Data Analytics. Esso offre una comprensione dei Big Data introducendo la teoria delle 3V, il modello delle 5V e la distinzione tra Data Analytics e Data Analysis. Si accenna, inoltre, alle varie categorie di Data Analytics e, infine, si discute il ciclo di vita della Big Data Analytics attraverso l'analisi delle sue 9 fasi.
- Il Capitolo 2 è dedicato all'analisi del software di Business Intelligence Power BI, esaminandone l'architettura e sottolineando l'importanza di Power Query nel processo di pulizia dei dati. Viene, anche, trattato l'aspetto della Data Visualization, descrivendo i diversi tipi di visualizzazioni, i filtri e le misure DAX.
- Nel Capitolo 3 viene effettuata un'esplorazione dei diversi tipi di dati: strutturati, non strutturati, semistrutturati e metadati e viene messo in evidenza il loro ruolo fondamentale nell'analisi delle informazioni. Inoltre, si approfondisce il processo di Estrazione, Trasformazione e Caricamento (ETL), presentando un resoconto dettagliato delle operazioni di ETL eseguite sul Dataset.
- Nel Capitolo 4 si descrivono le analisi generali effettuate sul Dataset, in particolare con tre report: il primo sui marcatori, il secondo sulle vittorie ai rigori e il terzo sulle partite giocate da ogni paese.
- Nel Capitolo 5 si rappresentano, invece, le analisi specifiche, con tre dashboard, riguardanti la tabella dei risultati, quella dei marcatori e l'ultima sulle nazionali dal 2000 al 2023.

CAPITOLO 1

Introduzione alla Data Analytics

Questo capitolo intende fornire una panoramica completa sul campo della Data Analytics. Si inizierà delineando le caratteristiche principali dei Big Data, esaminando la teoria delle 3V formulata da Doug Laney, per poi passare al più complesso modello delle 5V. Seguirà un'analisi dettagliata della Data Analytics, evidenziando le sue differenze rispetto alla Data Analysis e soffermandosi, in maniera particolare, sulle 4 categorie di Data Analytics. Concluderemo con una disamina approfondita del ciclo di vita della Big Data Analytics, esplorando ognuna delle nove fasi che lo caratterizzano.

1.1 Big Data

I Big Data, oggigiorno, hanno pervaso tutti i settori (industriale, commerciale ecc.), e stanno crescendo a un ritmo considerevole. Sono destinati a diventare un elemento cruciale nella competizione economica, equiparabili per importanza a risorse umane e capitale. I Big Data sono fondamentali per la crescita e l'innovazione; se gestiti e analizzati efficacemente, possono ottimizzare significativamente la produttività in vari ambiti, come quelli della pubblica amministrazione, dell'industria elettronica e informatica e della finanza. Grazie agli sviluppi raggiunti nel settore informatico e tecnologico, le aziende, oggi, adottano tecniche di Big Data Analytics per raccogliere e analizzare una vasta quantità di dati, allo scopo di trarne vantaggio. Questo processo è essenziale per raggiungere diversi obiettivi strategici, tra cui:

- l'aumento delle vendite;
- la scoperta di nuovi settori di mercato;
- l'incremento dell'interazione con i clienti;
- l'ottimizzazione delle attività operative;
- previsioni accurate.

I Big Data vanno aldilà della semplice Data Analysis e rappresentano una soluzione all'avanguardia con caratteristiche uniche e singolari. L'approccio evolve da una metodologia puramente statistica verso una più vasta basata su risorse interdisciplinari, integrando matematica, statistica e, soprattutto, i progressi nel campo informatico. Nonostante il crescente interesse per i Big Data e le tecniche di Data Analytics, molte aziende continuano a incontrare difficoltà nell'implementarli efficacemente. Infatti si stima che molte aziende, a livello

globale, non siano in grado di ottenere spunti significativi dai loro Big Data. In molte circostanze, il problema consiste nella difficoltà delle aziende di estrarre dati rilevanti prima che questi siano superati. In altri contesti, invece, l'ostacolo è rappresentato dalla carenza di informazioni.

Nella Figura 1.1 viene fornita una rappresentazione degli elementi che costituiscono il settore dei Big Data.

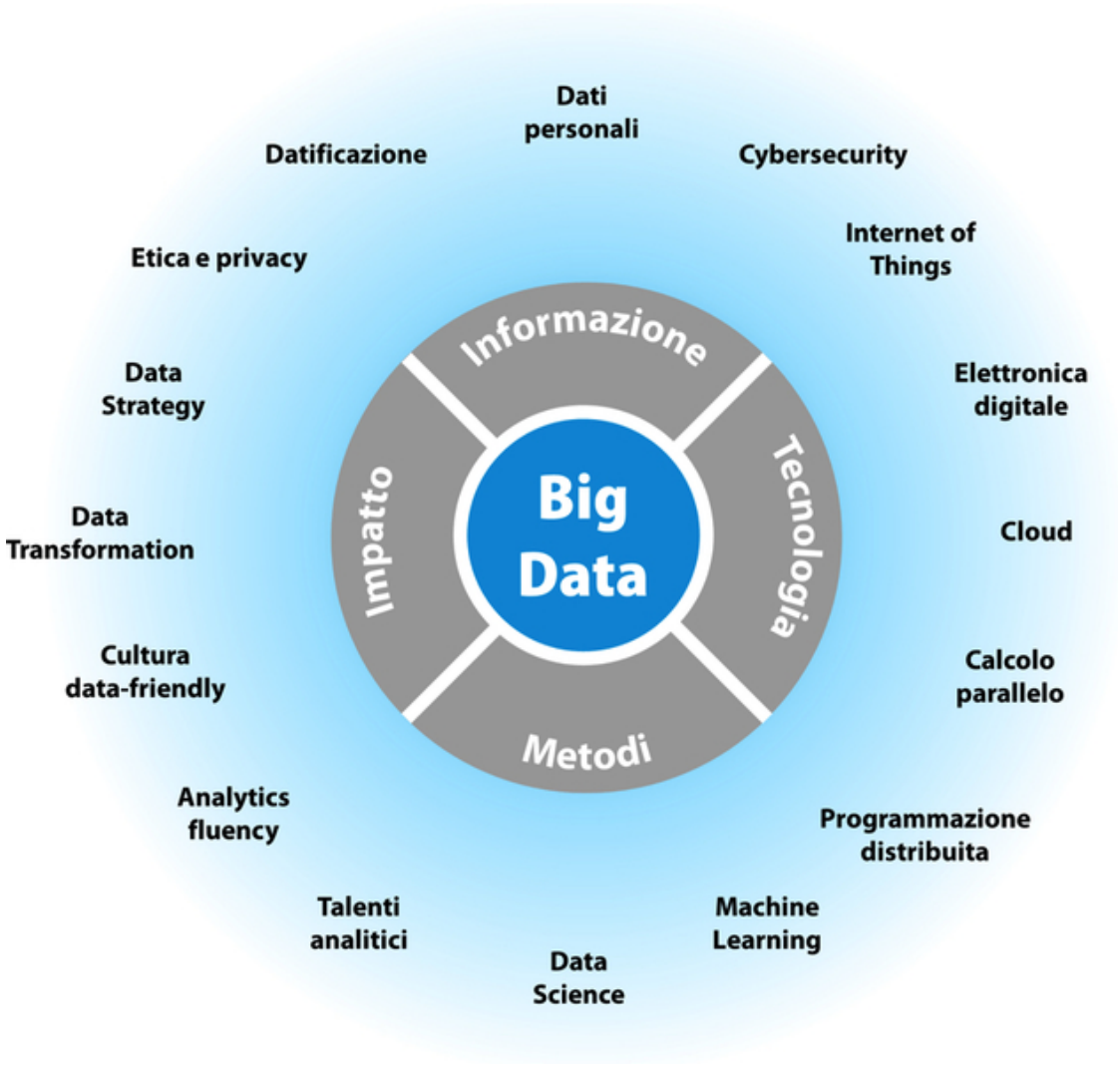


Figura 1.1: Elementi che costituiscono il settore dei Big Data

1.2 Le 5 V dei Big Data

Il concetto di "Big Data" si riferisce all'attività di raccogliere, memorizzare e analizzare grandi volumi di dati. Questo processo è caratterizzato dalle cosiddette 3V: Volume, Varietà e Velocità, termini che furono definiti per la prima volta nel 2001 da Doug Laney in un report per l'azienda Meta Group. Con il passare del tempo, si sono aggiunti altri due attributi importanti, ossia la Veracità e il Valore, che completano il quadro delle caratteristiche essenziali dei Big Data (Figura 1.2).

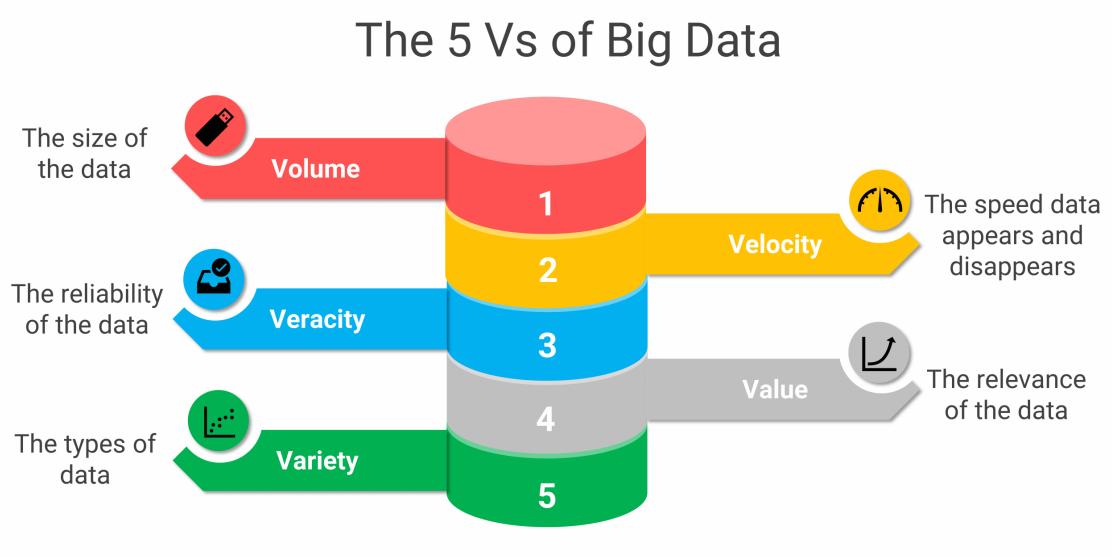


Figura 1.2: Le 5 V dei Big Data

1.2.1 Volume

Il quantitativo di dati trattati dalle soluzioni basate sui Big Data è rilevante e in continua espansione. La gestione di tali volumi implica requisiti complessi per l'archiviazione e l'elaborazione dei dati, nonché processi addizionali per la loro preparazione, manutenzione e controllo. Le organizzazioni e gli utenti generano quotidianamente circa 2,5 Exabyte (EB) di dati. Il volume delle informazioni può estendersi fino a raggiungere petabyte, o addirittura exabyte, necessitando, così, di architetture di memorizzazione per garantire una gestione efficiente dei dati.

1.2.2 Velocità

La Velocità dei Big Data descrive la rapidità con cui i dati vengono generati e distribuiti, accumulandosi in vasti insiemi informativi in tempi estremamente brevi. Un esempio è rappresentato dagli oltre 3,5 miliardi di ricerche giornaliere effettuate su Google che generano tantissimi GB di dati. È essenziale che le informazioni siano elaborate in tempo reale per influenzare immediatamente la vita delle persone e l'operatività delle aziende. Per garantire che i dati siano processati entro un lasso di tempo accettabile, le aziende devono implementare sistemi di elaborazione dei dati flessibili e devono possedere una significativa capacità di archiviazione.

1.2.3 Varietà

La Varietà dei dati si riferisce alla necessità di gestire molteplici formati e tipologie di dati attraverso le soluzioni di Big Data. Questa diversità presenta complesse sfide per le aziende, che devono affrontare questioni legate all'integrazione, alla trasformazione, all'elaborazione e alla conservazione delle informazioni. Ci sono diverse categorie di dati, tra cui i dati strutturati, che includono le informazioni su transazioni finanziarie o dati medici, i dati semi-strutturati, come quelli presenti nelle e-mail, e i dati non strutturati, nella forma di immagini.

1.2.4 Veracità

La Veracità indica la qualità e l'affidabilità dei dati. È essenziale che i dati introdotti negli ambienti di Big Data siano di alta qualità e "puliti"; pertanto, è indispensabile un'azione di data processing per eliminare i dati non validi e il rumore. I dati si classificano in due categorie: segnale e rumore. Il rumore comprende quei dati che non si possono trasformare in informazioni utili, e quindi non possiedono valore aggiunto. Il segnale, invece, si riferisce a dati che possono essere convertiti in informazioni rilevanti. Un alto rapporto segnale-rumore indica una veracità più alta. Generalmente, i dati raccolti in modo controllato presentano un livello inferiore di rumore, il che evidenzia come la qualità del rumore sia influenzata non solo dalla natura dei dati ma anche dalla loro fonte.

1.2.5 Valore

Il Valore dei Big Data risiede nella loro capacità di fornire informazioni preziose per le aziende e le organizzazioni. Questi dati vengono raccolti e analizzati allo scopo di generare informazioni per incrementare l'efficienza operativa, minimizzare i costi e perfezionare i processi. Un aspetto fondamentale del valore è legato alla veracità dei dati: quanto più sono attendibili, tanto maggiore sarà il loro valore per l'impresa. Nonostante ciò, il valore dei dati non dipende esclusivamente dalla loro affidabilità, ma anche dalla loro tempestività: essi tendono a perdere rilevanza con il trascorrere del tempo, analogamente a quanto accade a un prodotto che si avvicina alla propria scadenza. Pertanto, è cruciale assicurare un'elaborazione dei dati rapida, poiché esiste una relazione inversa tra tempo e valore: maggiore è il ritardo nell'analisi dei dati e minore sarà la loro utilità per l'azienda.

1.3 La Data Analytics

Il termine "Data Analytics" indica la scienza che gestisce l'intero ciclo di vita dei dati, un processo che comprende la raccolta, la pulizia, l'organizzazione, l'archiviazione e l'analisi degli stessi. Questa disciplina include anche lo sviluppo di metodi di analisi, tecniche scientifiche avanzate e l'impiego di tool automatici. Si concentra su metodologie che facilitano la Data Analysis attraverso l'impiego di tecnologie distribuite e fortemente scalabili; queste sono in grado di processare enormi quantità di informazioni provenienti da varie fonti. Il ciclo di vita dei Big Data implica l'identificazione, la raccolta, l'elaborazione e l'analisi di grandi volumi di dati grezzi e non elaborati. L'obiettivo è recuperare informazioni preziose che possano successivamente contribuire a identificare pattern o migliorare i dati già esistenti. Gli strumenti di Data Analytics trovano applicazione in molteplici settori. In particolare:

- nel campo scientifico servono a determinare le cause dei vari fenomeni e a migliorare l'accuratezza delle previsioni;
- nel contesto aziendale sono impiegati per ridurre i costi operativi e supportare decisioni strategiche;
- in ambienti basati sui servizi contribuiscono a focalizzare lo sviluppo di servizi di alta qualità a costi contenuti.

1.4 Le differenze tra Data Analytics e Data Analysis

La Data Analytics è un campo estremamente vasto che copre l'intero ciclo di vita dei dati. La Data Analysis, invece, si focalizza specificatamente sull'esame dei dati con l'intento

di trasformarli in informazioni rilevanti e impiegabili nelle decisioni strategiche. L'obiettivo principale è l'identificazione di trend, pattern, relazioni e previsione future. Pertanto, pur essendo solo una componente, la Data Analysis rappresenta un elemento cruciale all'interno del più ampio ambito della Data Analytics.

1.5 Categorie di Data Analytics

Esistono quattro categorie principali (Figura 1.3) che caratterizzano i diversi approcci nel settore della Data Analytics, ciascuna fondata su specifiche tecniche e algoritmi di analisi. Questa varietà implica che possano emergere differenti requisiti relativi ai dati, alla memorizzazione e all'elaborazione, conducendo a diverse tipologie di risultati. Nelle sottosezioni seguenti, forniremo un'introduzione a questi quattro tipi di Data Analytics.

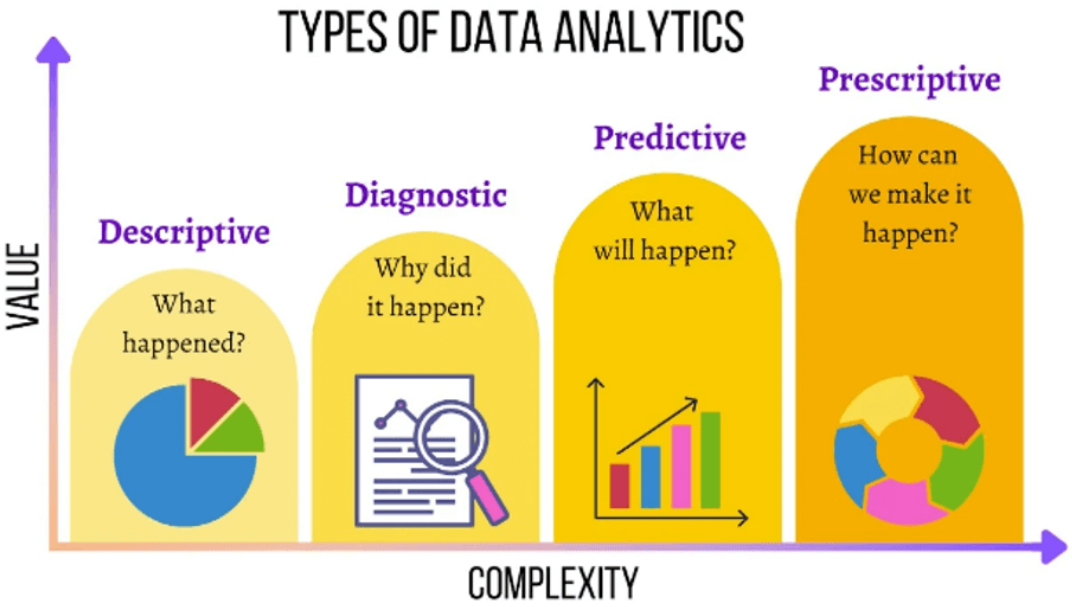


Figura 1.3: Categorie di Data Analytics

1.5.1 Analisi Descrittiva

L'analisi descrittiva ha come finalità quella di fornire risposte relative a eventi già accaduti; questa metodologia di analisi utilizza i dati per produrre informazioni. Si stima che l'80% dei risultati ottenuti attraverso la Data Analytics siano di natura descrittiva. In termini di valore aggiunto, questo tipo di analisi offre risultati meno significativi e si avvale di skill relativamente basilari per essere realizzata. Di solito si adottano sistemi di dashboard o di reporting; i report tendono a essere statici e presentano dati storici in vari formati. Generalmente, le query vengono eseguite su depositi di dati operazionali interni all'organizzazione.

1.5.2 Analisi Diagnostica

L'obiettivo dell'analisi diagnostica è identificare le origini di un evento passato, indagando le cause sottostanti che hanno dato luogo a tale fenomeno. Questo processo implica la ricerca di informazioni pertinenti che possano fornire risposte alle domande fondamentali

riguardanti le cause dell'evento, allo scopo di comprendere le ragioni per cui si è verificato. L'analisi diagnostica implica tipicamente la raccolta di dati da diverse fonti e la loro organizzazione in un formato adatto a effettuare analisi di tipo drill-down e roll-up. I risultati si possono vedere attraverso tool di visualizzazione interattiva che permettono agli utenti di rilevare trend e pattern. Le query effettuate in questo tipo di analisi sono generalmente più complesse rispetto a quelle usate nell'analisi descrittiva e vengono eseguite su dati multi-dimensionali conservati su piattaforme di elaborazione analitica avanzata.

1.5.3 Analisi Predittiva

L'analisi predittiva viene effettuata per prevedere gli esiti di eventi futuri. Questa tipologia di analisi arricchisce le informazioni disponibili, consentendo di avere maggiori conoscenze e di comprendere come queste siano relazionate agli eventi. La forza e l'intensità di queste relazioni costituiscono la base per i modelli predittivi, i quali si avvalgono degli eventi passati per formulare previsioni sul futuro. È fondamentale riconoscere che i modelli di analisi predittiva dipendono dalle condizioni specifiche in cui gli eventi passati si sono verificati. Qualora queste condizioni subiscano modifiche, è necessario aggiornare i modelli predittivi per mantenere la loro efficacia. Questa forma di analisi implica l'impiego di ampi dataset, che includono dati sia interni che esterni, e l'uso di metodologie avanzate di elaborazione dei dati; essa offre benefici superiori e necessita di skill più sofisticate rispetto all'analisi descrittiva e diagnostica.

1.5.4 Analisi Prescrittiva

L'analisi prescrittiva si basa sui risultati dell'analisi predittiva e mira a determinare le azioni più opportune da adottare. Questo tipo di analisi non si limita a identificare la migliore strategia possibile, ma esplora anche le ragioni dietro tale scelta. In sostanza, l'analisi prescrittiva fornisce risultati che permettono di effettuare valutazioni approfondite, poiché incorpora elementi che aiutano a interpretare il contesto sottostante. Di conseguenza, essa si rivela uno strumento efficace per ottenere dei vantaggi o per ridurre potenziali rischi. L'analisi prescrittiva si distingue come la più sofisticata tra le varie forme di Data Analytics, richiedendo skill più avanzate, nonché software e tool specializzati. Questo modello di analisi integra dati provenienti sia dall'interno che dall'esterno. I dati interni possono comprendere informazioni sulle vendite attuali e passate, sui prodotti e sulla soddisfazione dei clienti; i dati esterni, invece, possono includere informazioni provenienti da previsioni meteorologiche, social media e dati demografici.

1.6 Il ciclo di vita della Big Data Analytics

La Big Data Analytics si distingue dall'analisi dei dati tradizionale principalmente per le differenze nel volume, nella velocità e nella varietà delle informazioni che richiedono elaborazione. Pertanto, è essenziale adottare una metodologia strutturata, avanzando passo dopo passo nell'organizzazione delle attività. Il ciclo di vita della Big Data Analytics è articolato in diverse fasi, precisamente nove (Figura 1.5), che esamineremo dettagliatamente nelle sezioni seguenti.

1.6.1 Business Case Evaluation

Il ciclo di vita della Big Data Analytics inizia con l'identificazione di un business case ben definito, da cui vengono estratti le motivazioni e gli obiettivi specifici dell'analisi. Ini-

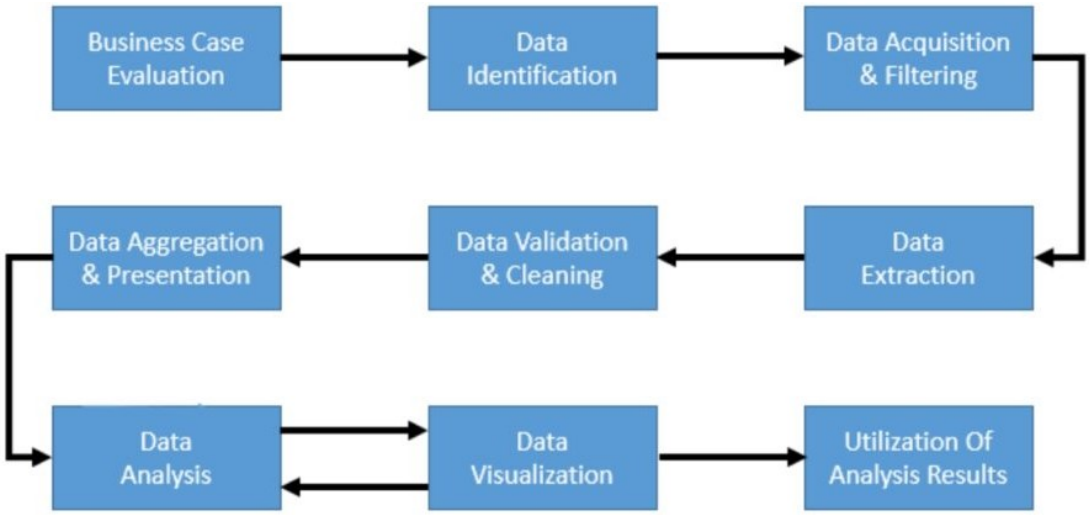


Figura 1.4: Ciclo di vita della Big Data Analytics

zialmente, è essenziale creare, validare e approvare il business case. Questa fase è di fondamentale importanza per i decision maker, i quali devono riconoscere le risorse di business e le sfide da superare prima di procedere con i task di analisi effettivi. Nel corso della revisione del business case, è possibile determinare i KPI (Key Performance Indicator), che sono gli indicatori impiegati per misurare il successo di un'attività. Se i KPI non sono immediatamente disponibili, è necessario formulare obiettivi che siano SMART; questo acronimo sta ad indicare che devono essere Specifici, Misurabili, Raggiungibili, Rilevanti e Tempestivi. Inoltre, durante questa fase, si determina se la situazione rappresenta un problema di Big Data, valutando le cinque caratteristiche fondamentali dei Big Data: Volume, Varietà, Velocità, Veracità e Valore. Un altro scopo cruciale di questo stadio è stabilire il budget richiesto per il progetto, che deve includere tutte le spese relative all'acquisizione di hardware, di strumenti e alla formazione del personale.

1.6.2 Data Identification

La fase di Data Identification mira a riconoscere i dataset necessari per un progetto di analisi e a identificare una gamma di fonti di dati, al fine di avere più possibilità di scoprire pattern. I dataset possono derivare sia da fonti interne che esterne all'impresa. Per i dataset interni, si procede alla compilazione di un elenco di dataset già disponibili tramite fonti interne, inclusi i data mart e i sistemi operazionali; tale elenco è poi confrontato con una specifica predefinita di dataset. Per quanto riguarda i dataset esterni, si compila un elenco di potenziali data provider di terze parti, che include data market e dataset aperti al pubblico.

1.6.3 Data Acquisition and Filtering

Nella fase di Data Acquisition and Filtering si procede alla raccolta dei dati dalle sorgenti precedentemente individuate. Questi dati vengono successivamente filtrati per rimuovere quelli privi di valore o corrotti, dove per "corrotti" si intendono record con valori nulli, senza senso o non validi. È essenziale effettuare un backup del dataset originale prima di procedere al filtraggio, poiché potrebbe rivelarsi prezioso per analisi future. Inoltre, è necessario assicurarsi che i dati acquisiti siano resi persistenti, sia prima sia dopo l'analisi, a seconda che si utilizzi analitica batch o realtime. Il processo di acquisizione può essere ottimizzato tramite l'integrazione di metadati, sia per le fonti di dati interne che esterne. I metadati de-

vono essere redatti in un formato leggibile da macchina e devono essere elaborati attraverso i vari livelli di analisi.

1.6.4 Data Extraction

Spesso una parte delle informazioni si presenta in formati non compatibili con le soluzioni di Big Data, aumentando il rischio di dover trattare tipologie di dati eterogenei, soprattutto quando provengono da fonti esterne. Lo scopo di questa fase consiste nel raccogliere informazioni di varia natura e trasformarle in un formato appropriato per l'analisi dei dati. Naturalmente, i metodi di estrazione e conversione dipendono dal tipo di analisi prevista e dalle specifiche della soluzione di Big Data utilizzata.

1.6.5 Data Validation and Cleansing

È cruciale notare che dati inesatti possono gravemente compromettere la qualità dei risultati derivanti dalle analisi. Nelle operazioni aziendali tradizionali, i dati presentano una struttura definita e passano attraverso un processo di pre-convalida. Al contrario, nel contesto dei Big Data, si affronta spesso l'acquisizione di dati non strutturati, che possono mancare di standard di validità. La complessità dei dati può rendere difficile stabilire criteri di validazione adeguati. In questa situazione, diventa fondamentale l'implementazione del processo di Data Validation and Cleansing, il quale mira a stabilire regole rigorose per la validazione e a rimuovere i dati riconosciuti come non validi. Le tecnologie basate sui Big Data spesso raccolgono dati ridondanti da vari dataset. Tale ridondanza può essere sfruttata strategicamente per esplorare le relazioni tra i dataset al fine di sviluppare criteri di validazione accurati e per integrare dati validi mancanti. Per l'analitica batch, la convalida e la pulizia dei dati possono avvenire tramite procedure di ETL offline. Per l'analitica real time, è essenziale un sistema in-memory più sofisticato capace di esaminare e correggere i dati al momento del loro arrivo dalla fonte. L'origine dei dati può influenzare significativamente l'accuratezza e la qualità delle informazioni; tuttavia, dati inizialmente considerati non validi possono rivelarsi preziosi come indicatori di pattern e trend nascosti.

1.6.6 Data Aggregation and Representation

I dati possono essere organizzati in molteplici dataset; è, quindi, necessario che siano collegati tramite elementi comuni, come l'identificatore o la data. In alcuni casi gli stessi attributi li ritroviamo in più insiemi di dati. Pertanto, è cruciale implementare un sistema che permetta la riconciliazione dei dati o determinare quale dataset rappresenta le informazioni più appropriate. Il processo di Data Aggregation and Representation si concentra sull'aggregazione di vari dataset per fornire una prospettiva integrata. Questa operazione può rivelarsi complessa a causa delle discrepanze nelle strutture dati e nelle interpretazioni semantiche, dove termini diversi in insiemi di dati distinti potrebbero indicare lo stesso concetto. L'elaborazione di volumi ingenti di dati tramite tecnologie di Big Data può rendere l'aggregazione un processo dispendioso in termini di tempo e risorse. Superare queste discrepanze può necessitare di algoritmi sofisticati che operano automaticamente, eliminando la necessità di interventi umani. Durante questa fase è fondamentale tenere in considerazione le future richieste di analisi di dati per garantire che le informazioni siano riutilizzabili. Indipendentemente dalla necessità di aggregare i dati, è essenziale riconoscere che le informazioni possono essere memorizzate in forme diverse, ognuna delle quali potrebbe essere più adatta a seconda dell'analisi prevista.

1.6.7 Data Analysis

La fase di Data Analysis si concentra sull'elaborazione e sull'interpretazione delle informazioni accumulate. Durante questa fase, il procedimento può manifestare una natura ciclica, soprattutto se l'analisi persegue scopi esplorativi. In tale contesto, l'analisi viene ripetuta costantemente fino all'identificazione di modelli o correlazioni significative. La complessità di questo step dipende strettamente dagli scopi prefissati. In alcune situazioni, potrebbe bastare interrogare un insieme di dati per raccogliere e comparare informazioni. In contesti più complicati, invece, è necessario l'impiego di metodologie avanzate di data mining e analisi statistica per rilevare modelli e discrepanze o per creare modelli matematici o statistici che illustrano le interazioni tra le variabili. L'analisi dei dati si suddivide in due tipologie fondamentali: confermativa ed esplorativa.

L'analisi confermativa adotta un approccio deduttivo e stabilisce inizialmente una causa presunta del fenomeno studiato, detta ipotesi; questa viene, poi, analizzata per confermare o smentire la stessa. In seguito, le informazioni raccolte sono sottoposte a un'analisi accurata al fine di confermare o smentire le ipotesi iniziali e offrire soluzioni chiare a quesiti specifici. In questa situazione, si usano unicamente tecniche di campionamento dei dati mentre vengono tralasciati risultati atipici o anomalie.

L'analisi esplorativa adotta una prospettiva induttiva correlata al data mining, priva di ipotesi predefinite. Si analizzano i dati per comprendere la causa del fenomeno osservato. Sebbene questo approccio possa non fornire conclusioni definitive, offre un aiuto nella scoperta di pattern o anomalie nei dati esaminati.

1.6.8 Data Visualization

La capacità di analizzare ampie quantità di dati e individuare intuizioni di valore, rischia di essere inutile se solo gli analisti sono capaci di decifrare gli esiti dell'analisi. Il processo di Data Visualization mira a utilizzare tecniche e strumenti di rappresentazione grafica al fine di comunicare efficacemente le conclusioni dell'analisi, rendendo, così, più agevole per gli utenti aziendali l'interpretazione accurata dei dati. Questi utenti devono essere dotati delle competenze adeguate per assimilare tali dati, con lo scopo di beneficiare delle analisi e fornire a loro volta un feedback costruttivo. Concluso questo processo, gli utenti saranno in grado di svolgere autonomamente analisi visive, aprendo la possibilità di esplorare risposte a quesiti fino ad allora non considerati. Le informazioni emerse possono variare nel modo in cui vengono presentate, influenzando, di conseguenza, la loro interpretazione. È, quindi, essenziale selezionare metodi di visualizzazione appropriati, tenendo in considerazione il contesto aziendale specifico.

1.6.9 Utilization of Analysis Results

In questa fase si procede all'identificazione delle modalità e delle aree in cui applicare i risultati derivati dall'analisi dei dati. A seconda delle caratteristiche del problema analizzato, tale analisi può portare alla creazione di modelli che offrono nuove conoscenze e permettono di riconoscere schemi o relazioni tra diversi elementi. Questi modelli si manifestano comunemente come equazioni matematiche o insiemi di regole. Questi ultimi trovano impiego nel miglioramento della logica dei processi aziendali e dei sistemi applicativi, contribuendo significativamente all'ottimizzazione delle performance generali. I risultati dell'analisi possono essere integrati, sia automaticamente sia manualmente, nei sistemi di gestione aziendale per incrementarne la produttività e l'efficacia. Infine, la capacità di identificare pattern, correlazioni e anomalie può essere utilizzata per affinare e migliorare i processi aziendali. Ciò avviene attraverso interventi mirati basati sulle informazioni ottenute dall'analisi, rendendo

i risultati analitici uno strumento fondamentale per l'ottimizzazione e il miglioramento dei processi di business.

CAPITOLO 2

Introduzione a Power BI

Nel presente capitolo, intendiamo esaminare e approfondire il software di Business Intelligence, Power BI, uno strumento per l'analisi e la visualizzazione dei dati ampiamente adottato. Verrà analizzata l'architettura di Power BI, con un focus particolare sul componente chiave del sistema, ovvero Power BI Desktop. Inoltre, verrà discusso il processo di data cleaning all'interno di Power BI, esplorando le funzionalità di Power Query e la sua rilevanza nell'elaborazione dei dati. Infine, ci concentreremo sulla componente di Data Visualization del software, investigando i vari tipi di visualizzazione offerti, l'utilizzo dei filtri e l'implementazione delle misure DAX per migliorare e personalizzare l'analisi dei dati.

2.1 Power BI

Power BI è una piattaforma di Business Intelligence creata da Microsoft e progettata per supportare l'analisi dei dati aziendali mediante grafici interattivi e un'interfaccia intuitiva. Questo strumento permette agli utenti di generare con precisione e velocità report dettagliati e dashboard informative. Tale piattaforma offre un insieme di servizi software, applicazioni e connettori creati per trasformare dati provenienti da varie fonti in informazioni organizzate e visualmente intuitive. Le fonti di dati con cui Power BI può interfacciarsi includono, ad esempio, fogli di calcolo Excel, sistemi di data warehouse cloud-based e configurazioni ibride locale-cloud. Power BI supporta l'accesso a queste fonti, la visualizzazione delle informazioni pertinenti e la loro distribuzione agli stakeholder. Lanciato di recente, si è rapidamente affermato come uno strumento competitivo nel campo della Business Intelligence. La sua crescita è stata alimentata da significativi investimenti di Microsoft e dalla sua capacità di competere con leader di mercato come Tableau e QlikSense. L'aggiornamento costante, con nuove versioni rilasciate mensilmente, e l'accento posto sull'innovazione continua, hanno posizionato Power BI come leader nel Magic Quadrant di Gartner, per gli strumenti di analisi dei dati.

2.2 Architettura di Power BI

Power BI si articola attraverso diversi componenti interconnessi, come illustrato in Figura 2.1. Gli elementi fondamentali includono:

- *Power BI Desktop* : una piattaforma desktop per sistemi Windows progettata per la creazione, modifica e visualizzazione di report. Questa applicazione fornisce agli utenti strumenti per l'analisi e per la gestione dei dati.

- *Power BI Service*: un servizio online strutturato secondo il modello Software as a Service (SaaS), che permette la visualizzazione e la condivisione di dashboard aggiornate in tempo reale. Questa funzione promuove la collaborazione e il monitoraggio delle performance aziendali.
- *Gateway di Power BI*: i gateway facilitano la sincronizzazione del flusso dei dati in entrata e in uscita nel sistema Power BI, massimizzando l'efficienza e l'efficacia del processo.
- *Power BI Mobile*: disponibile per dispositivi iOS e Android, questa applicazione permette di monitorare e accedere ai dati elaborati attraverso Power BI Desktop da qualsiasi dispositivo mobile, offrendo, così, flessibilità e accessibilità alle informazioni aziendali.

In aggiunta ai quattro componenti principali, Power BI include anche altri due elementi essenziali, ovvero:

- *Power BI Report Server*: un server di reportistica locale che permette la pubblicazione di report sviluppati mediante Power BI Desktop.
- *Power BI Report Builder*: uno strumento specificamente progettato per la creazione di report impaginati, destinati alla condivisione attraverso il servizio Power BI.

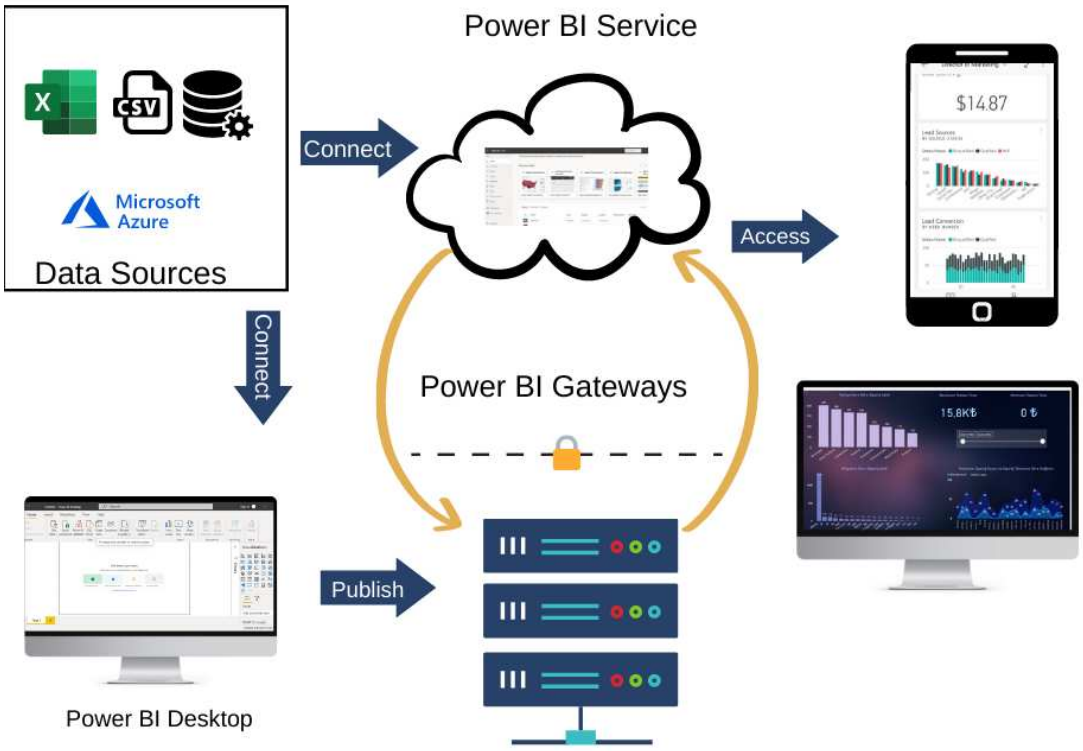


Figura 2.1: Architettura di Power BI

2.3 Power BI Desktop

Il nostro studio si concentrerà sull'utilizzo di Power BI Desktop, uno strumento avanzato che consente l'importazione di dati da una vasta gamma di sorgenti, permettendoci di elaborare ed estrarre informazioni utili.

Il processo operativo in Power BI generalmente include:

- Eseguire operazioni di data shaping tramite query per creare modelli di dati efficaci.
- Stabilire collegamenti con svariate fonti di dati.
- Utilizzare tali modelli per generare visualizzazioni e report
- Condividere i file dei report con altri utenti, consentendo loro di utilizzarli, ampliarli e distribuirli. La modalità più efficace consiste nel caricarli nel servizio Power BI.

Power BI Desktop integra la robusta tecnologia di Microsoft Query Engine con avanzate funzionalità di visualizzazione e modellazione dei dati. In questo modo, gli analisti dei dati e altri utenti possono agevolmente sviluppare e distribuire insiemi di query, report e modelli. La collaborazione tra Power BI Desktop e il servizio Power BI permette di gestire, elaborare, condividere ed espandere le informazioni dettagliate ricavate dai dati con maggiore facilità. In sintesi, Power BI Desktop semplifica e ottimizza il processo di progettazione e creazione di report e repository di Business Intelligence, che, altrimenti, risulterebbe disorganizzato e complesso.

La schermata iniziale di Power BI Desktop mette a disposizione tre tipologie di visualizzazioni:

- *Report*: permette agli utenti di utilizzare le query create per produrre rappresentazioni grafiche, organizzandole su una o più pagine.
- *Dati*: consente di esaminare le informazioni caricate nel report attraverso un modello di dati, al quale è possibile aggiungere misure, creare colonne aggiuntive e verificare le relazioni esistenti.
- *Relazioni*: permette di visualizzare, gestire e apportare modifiche a una rappresentazione grafica delle relazioni definite allinterno del modello di dati.

Le tre modalità di visualizzazione in Power BI possono essere attivate tramite le rispettive icone situate sul lato sinistro dell'interfaccia. Ad esempio nella Figura 2.2, è stata selezionata la modalità di Visualizzazione Report.

Figura 2.2: Modalità di visualizzazione in Power BI Desktop

2.4 Data Cleaning

Dopo aver completato l'installazione di Power BI Desktop, si avrà accesso a una vasta e crescente gamma di dati. Per esplorare le diverse fonti di dati disponibili, è necessario selezionare "Recupera dati" > "Altro" nella sezione Home di Power BI Desktop. Successivamente, nella finestra "Recupera dati", sarà possibile consultare l'elenco completo delle fonti di dati disponibili (Figura 2.3).

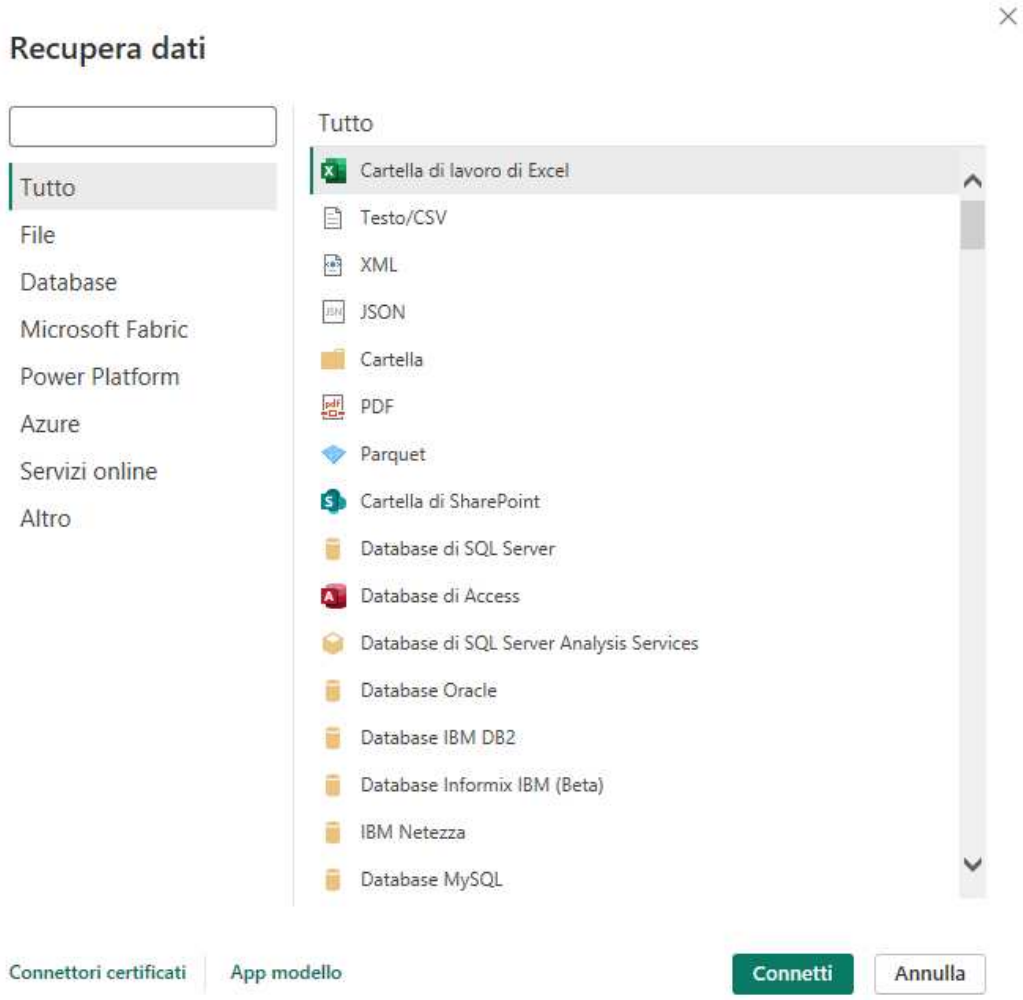
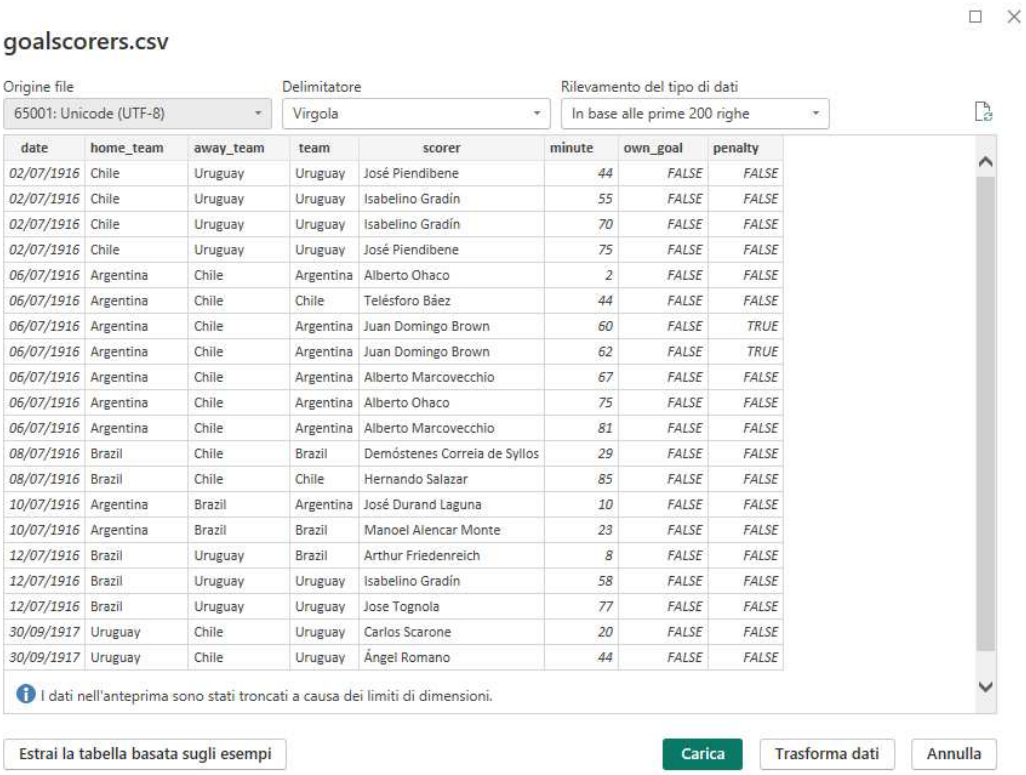


Figura 2.3: Finestra "Recupera dati" in Power BI Desktop

Una volta stabilito il collegamento con la fonte dei dati, si passa alla fase di trasformazione e pulizia degli stessi. Questa fase è di fondamentale importanza e incide significativamente sulla qualità delle analisi effettuate sui dati. Tra le operazioni comuni in questa fase troviamo la sostituzione dei valori, l'eliminazione delle duplicazioni, la rimozione dei dati errati e l'adeguamento dei tipi di dati.

2.4.1 Power Query Editor

Il software Power BI Desktop include il componente Power Query Editor, che si apre in una finestra separata. Attraverso l'Editor di Power Query, è possibile creare interrogazioni e manipolare i dati, per poi importare il modello dei suddetti perfezionato in Power BI Desktop per l'elaborazione dei report. Una volta stabilita la connessione, apparirà la finestra dello strumento di navigazione come illustrato nella Figura 2.4.



date	home_team	away_team	team	scorer	minute	own_goal	penalty
02/07/1916	Chile	Uruguay	Uruguay	José Piendibene	44	FALSE	FALSE
02/07/1916	Chile	Uruguay	Uruguay	Isabelino Gradín	55	FALSE	FALSE
02/07/1916	Chile	Uruguay	Uruguay	Isabelino Gradín	70	FALSE	FALSE
02/07/1916	Chile	Uruguay	Uruguay	José Piendibene	75	FALSE	FALSE
06/07/1916	Argentina	Chile	Argentina	Alberto Ohaco	2	FALSE	FALSE
06/07/1916	Argentina	Chile	Chile	Telésforo Bäez	44	FALSE	FALSE
06/07/1916	Argentina	Chile	Argentina	Juan Domingo Brown	60	FALSE	TRUE
06/07/1916	Argentina	Chile	Argentina	Juan Domingo Brown	62	FALSE	TRUE
06/07/1916	Argentina	Chile	Argentina	Alberto Marcovecchio	67	FALSE	FALSE
06/07/1916	Argentina	Chile	Argentina	Alberto Ohaco	75	FALSE	FALSE
06/07/1916	Argentina	Chile	Argentina	Alberto Marcovecchio	81	FALSE	FALSE
08/07/1916	Brazil	Chile	Brazil	Demóstenes Correia de Syllos	29	FALSE	FALSE
08/07/1916	Brazil	Chile	Chile	Hernando Salazar	85	FALSE	FALSE
10/07/1916	Argentina	Brazil	Argentina	José Durand Laguna	10	FALSE	FALSE
10/07/1916	Argentina	Brazil	Brazil	Manoel Alencar Monte	23	FALSE	FALSE
12/07/1916	Brazil	Uruguay	Brazil	Arthur Friedenreich	8	FALSE	FALSE
12/07/1916	Brazil	Uruguay	Uruguay	Isabelino Gradín	58	FALSE	FALSE
12/07/1916	Brazil	Uruguay	Uruguay	Jose Tognola	77	FALSE	FALSE
30/09/1917	Uruguay	Chile	Uruguay	Carlos Scarone	20	FALSE	FALSE
30/09/1917	Uruguay	Chile	Uruguay	Ángel Romano	44	FALSE	FALSE

Figura 2.4: Finestra "strumento di navigazione" in Power BI Desktop

In questa sezione, è possibile esaminare un'anteprima dei dati; si ha l'opportunità di scegliere tra "Carica", per importare direttamente la tabella, o "Trasforma dati", per apportare modifiche prima del caricamento. Selezionando "Trasforma dati", l'editor di Power Query verrà avviato; avremo, così, una vista rappresentativa della tabella. Nel lato destro della finestra, è situato il pannello "Impostazioni query", come illustrato nella Figura 2.5, accessibile anche dalla scheda "Visualizza" all'interno dell'editor di Power Query. A questo punto, è possibile modificare i dati secondo le proprie esigenze, ovvero eseguire un data shaping. Durante l'importazione e la visualizzazione, per modificare i dati è necessario fornire precise direttive all'editor di Power Query. Questo processo non comporterà alcuna alterazione della fonte dei dati originale, ma modificherà unicamente questa vista specifica. Il data shaping comporta trasformazioni dei dati, come la rinominazione di colonne o tabelle, l'eliminazione di righe o colonne, o la modifica dei tipi di dati. L'Editor di Power Query documenta tali operazioni in sequenza nell'elenco "Passaggi applicati" situato nel riquadro delle Impostazioni Query. Quando la query si connette all'origine dei dati, questi passaggi vengono eseguiti sistematicamente, assicurando che i dati siano sempre conformi alla struttura definita.

2.5 Data Visualization

La Data Visualization si riferisce all'impiego di elementi visivi, quali diagrammi, grafici e, occasionalmente, animazioni, con l'obiettivo di rappresentare informazioni derivate dai dati. Questo approccio permette di comunicare in modo chiaro e immediato le complesse relazioni tra i dati e le conoscenze correlate. Power BI facilita una visualizzazione dei dati veloce ed efficace, mettendo a disposizione una vasta gamma di strumenti per potenziare questa fase. Tra tali strumenti troviamo:

- una varietà di tipi di visualizzazione;

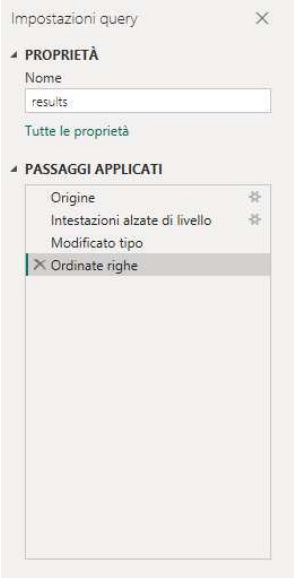


Figura 2.5: Pannello "impostazioni query" in Power BI Desktop

- filtri;
- Data Analysis eXpressions.

Nelle prossime sezioni, approfondiremo ciascuno di questi elementi.

2.5.1 Tipi di visualizzazione

Una volta sviluppato un modello di dati, si può procedere all'inserimento dei vari campi nell'area di disegno della vista report, generando rappresentazioni grafiche delle informazioni contenute nel modello stesso. Queste sono comunemente note come oggetti visivi. In determinate circostanze, potrebbe essere necessario creare un insieme di oggetti visivi, destinati a fornire un'analisi dettagliata dei dati utilizzati per la costruzione del modello tramite Power BI. Questo insieme è denominato report; la Figura 2.6 illustra un esempio delle diverse visualizzazioni disponibili in Power BI. Vedremo di seguito alcuni esempi:

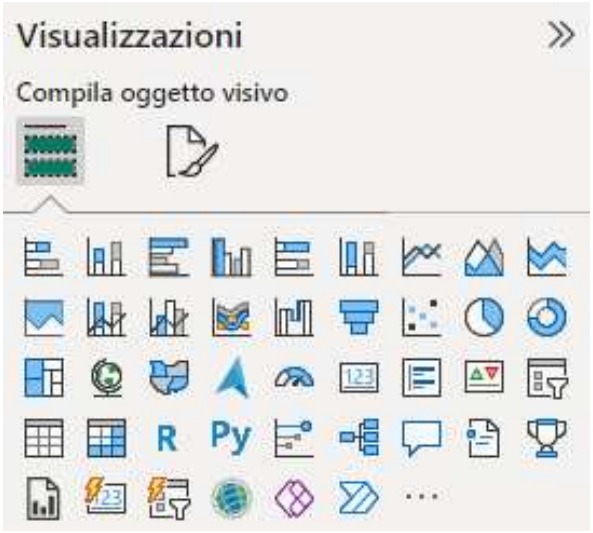


Figura 2.6: Pannello "visualizzazioni" in Power BI Desktop

- *Grafici a torta e a barre sovrapposte*: tali grafici sono suddivisi in sezioni che rappresentano le porzioni di un insieme. Essi forniscono un metodo efficace per strutturare i dati e confrontare le dimensioni relative di ciascun componente.
- *Grafici a linee e ad aree*: questi grafici rappresentano la variazione di una o più quantità nel tempo mediante il tracciamento di una sequenza di data point. Sono frequentemente impiegati nell'analisi predittiva. I grafici a linee illustrano tali variazioni attraverso linee connesse tra i data point, mentre i grafici ad area collegano i punti con segmenti di linea e utilizzano il colore per sovrapporre e distinguere le variabili rappresentate.
- *Istogrammi*: questo tipo di rappresentazione grafica mostra la distribuzione di valori numerici attraverso un diagramma a barre contigue che indicano la quantità di dati inclusi in specifici intervalli. Questa forma di rappresentazione consente all'osservatore di individuare facilmente eventuali anomalie presenti all'interno di un insieme di dati.
- *Tabelle*: composte da righe e colonne, vengono utilizzate per il confronto tra variabili. Le tabelle possono presentare un'ampia quantità di informazioni in modo organizzato; tuttavia, possono risultare eccessivamente complesse per gli utenti che cercano solo di individuare tendenze generali.
- *Schede con numero singolo*: queste schede visualizzano un singolo dato, ossia un'informazione numerica isolata. In alcune circostanze, l'obiettivo principale di un report di Power BI è presentare un solo valore come, ad esempio, il totale delle partite.
- *Schede con più righe*: queste schede presentano molteplici elementi informativi con ciascun valore numerico disposto su righe separate, consentendo così una visualizzazione chiara e ordinata dei data point.

2.5.2 Filtri

Una delle funzionalità principali di Power BI Desktop è la capacità di applicare filtri ai dati. Durante la creazione di un report, è possibile utilizzare vari tipi di filtri per ottimizzare la visualizzazione grafica delle informazioni. Nel pannello "Filtri" possiamo individuare tre categorie che ci consentono di analizzare dettagliatamente le visualizzazioni:

- *Filtro per pagina*: applicato a tutti gli elementi visivi presenti nella pagina del report.
- *Filtro per elemento visivo*: applicato a un singolo elemento visivo all'interno di una pagina del report.
- *Filtro per lintero report*: si applica a tutte le pagine del report.

Il pannello "Filtri" offre la possibilità di utilizzare campi non ancora inclusi nei vari elementi grafici del report. Ogni filtro può essere ulteriormente suddiviso in sotto-filtri, che permettono di affinare l'analisi dei dati. Tra i principali tipi di sotto-filtri possiamo identificare:

- *Filtro avanzato*: consente di filtrare i dati applicando condizioni specifiche (ad esempio: maggiore di, minore di, uguale a, etc.).
- *Filtro basilare*: ideale per selezionare valori specifici (ad esempio: anno, mese, giorno).
- *Filtro Top N*: utile per analizzare dei valori specifici.

2.5.3 Data Analysis eXpressions

Nel corso dell'analisi dei dati in Power BI vengono effettuati calcoli personalizzati, noti come misure, indispensabili per esaminare i dati in modo aggregato e fornire informazioni significative. Queste misure possono essere integrate nelle visualizzazioni di Power BI come elementi di dati all'interno di grafici, tabelle o altre rappresentazioni; esse si basano su risultati ottenuti mediante specifiche espressioni matematiche.

Durante la creazione di misure personalizzate, si impiega il linguaggio DAX (Data Analysis eXpressions), rinomato per la sua estesa libreria che comprende oltre 200 funzioni, operatori e strutture. Tale libreria offre una notevole flessibilità nella generazione delle misure e nell'elaborazione delle espressioni di calcolo. Le formule DAX presentano somiglianze con le formule di Excel, condividendo molte funzioni comuni, come SUM, DATE e LEFT; nonostante ciò, le funzioni DAX sono specificamente progettate per essere utilizzate in un contesto di dati relazionale, tipico di Power BI Desktop. Si evince che l'uso di DAX è particolarmente adatto per la gestione e l'analisi delle informazioni su questa piattaforma.

Durante l'elaborazione delle formule tramite DAX, è di cruciale importanza considerare attentamente il tipo di dato per evitare incongruenze e incoerenze nei risultati ottenuti. I tipi di dati compatibili comprendono numeri interi, decimali, valute, booleani, testo e date. Nelle espressioni DAX, inoltre, è essenziale specificare correttamente tabelle e colonne, utilizzando gli apici nel caso in cui il nome della tabella contenga spazi. Infine, per racchiudere colonne e misure, è necessario utilizzare le parentesi quadre.

CAPITOLO 3

Descrizione del dataset ed attività di ETL

In questo capitolo esamineremo i diversi tipi di dati: strutturati, non strutturati, semi-strutturati e metadati, sottolineando il loro ruolo essenziale nell'analisi delle informazioni. Ci concentreremo sul dataset riguardante i risultati delle partite di calcio delle squadre nazionali, per poi approfondire il processo di Estrazione, Trasformazione e Caricamento (Extraction, Transformation and Loading-ETL). Infine, presenteremo un resoconto dettagliato delle operazioni di ETL eseguite sul nostro dataset utilizzando il software Power BI.

3.1 Introduzione sui tipi di dati

Alla base di una campagna di Data Analytics vi sono i dati, i quali possono derivare da diverse fonti e presentarsi in vari formati. In particolare, i dati elaborati tramite soluzioni Big Data possono essere generati da esseri umani o da macchine. Più nello specifico:

- I dati generati dall'uomo sono il risultato dell'interazione tra esseri umani e sistemi. Alcuni esempi sono dati provenienti da e-mail, social media e sistemi di condivisione di foto.
- I dati generati dalle macchine sono prodotti da software e dispositivi hardware in risposta a eventi del mondo reale.

Come osservato, sia i dati generati dall'uomo che quelli generati dalle macchine provengono da molteplici fonti eterogenee e possono variare nel formato e nel tipo. Le tre principali categorie di dati processati nelle soluzioni di Big Data sono:

- *Dati strutturati*;
- *Dati non strutturati*;
- *Dati semi-strutturati*.

Queste tre categorie si riferiscono alla loro struttura interna e, talvolta, sono definite come formati dei dati. Oltre a questi tipi, nell'ambito dei Big Data, assumono particolare importanza anche i metadati.

3.1.1 Dati strutturati

I dati strutturati si riferiscono a informazioni che seguono un modello o schema specifico. Tali informazioni sono solitamente organizzate in tabelle, il che consente di mettere in evidenza le relazioni tra diverse entità; per questo motivo, vengono spesso memorizzate in un database relazionale. Questo tipo di dati è comunemente generato da sistemi informativi, come i sistemi di pianificazione delle risorse aziendali (ERP) e i sistemi di gestione delle relazioni con i clienti (CRM). Grazie alla vasta disponibilità di tool e database che gestiscono nativamente i dati strutturati, non è necessario adottare particolari precauzioni per la loro elaborazione o conservazione.

3.1.2 Dati non strutturati

I dati non strutturati rappresentano informazioni che non aderiscono a uno schema o ad un modello di dati specifico. È stato stimato che, in una tipica organizzazione, fino all'80% dei dati rientrano in questa categoria. La quantità di dati non strutturati tende ad aumentare a un ritmo più rapido rispetto ai dati strutturati. Queste informazioni possono essere di natura testuale o binaria e sono spesso archiviate in file indipendenti che non hanno relazioni intrinseche tra loro. Un file di testo può contenere materiale proveniente da varie fonti, come articoli di blog e tweet. I file binari, invece, sono generalmente file multimediali che contengono informazioni sotto forma di immagini, suoni o video. Sebbene sia i file di testo che quelli binari abbiano una struttura determinata dal formato del file, la caratteristica distintiva dei dati non strutturati riguarda il corrispettivo formato, piuttosto che la struttura del file stesso. I dati non strutturati non possono essere direttamente elaborati o interrogati utilizzando SQL. Nel caso sia necessario conservare tali dati allinterno di un database relazionale, gli stessi vengono memorizzati in una tabella strutturata come BLOB (Binary Large Object). In alternativa, esistono i database Not-only-SQL (NoSQL), una tipologia di database non relazionali, particolarmente adatti per l'archiviazione di dati non strutturati. Questi ultimi sono in grado di memorizzare dati non strutturati insieme a dati strutturati.

3.1.3 Dati semi-strutturati

I dati semi-strutturati possiedono una forma strutturale e una coerenza specifiche; tuttavia, non rientrano nella categoria dei dati relazionali. Questi dati, al contrario, conservano una natura gerarchica o basata su modelli a grafo. Tali informazioni vengono principalmente archiviate in documenti testuali; esempi comuni di rappresentazioni di dati semi-strutturati includono file XML e JSON. La loro natura testuale e la conformità a una struttura definita rendono più agevole la loro elaborazione rispetto a quella dei dati completamente non strutturati. Fonti tipiche di dati semi-strutturati includono file EDI (Electronic Data Interchange), feed RSS, fogli di calcolo e dati provenienti da sensori. La gestione di questi dati richiede spesso particolari accorgimenti riguardo alla loro pre-elaborazione e archiviazione, specialmente quando il loro formato non è testuale. Un esempio specifico di pre-elaborazione dei dati semi-strutturati è la validazione di un file XML.

3.1.4 Metadati

I metadati rivestono un ruolo cruciale nell'attuale panorama dei dati; essi forniscono una descrizione dettagliata delle caratteristiche e della struttura di un dataset. Una considerevole parte di queste informazioni, generate da dispositivi e software automatici, viene comunemente associata ai dati a cui si riferiscono. È di primaria importanza tracciare e gestire con precisione i metadati durante l'elaborazione, l'archiviazione e l'analisi dei Big Data.

Tutto ciò deriva dal fatto che i metadati forniscono elementi essenziali riguardanti l'origine e la genealogia dei dati, risultando particolarmente utili per comprendere il contesto e la validità delle informazioni nel corso del trattamento. Le soluzioni basate sui Big Data fanno grande affidamento sui metadati, in particolare quando si tratta di elaborare informazioni semi-strutturate o non strutturate. Inoltre, è importante sottolineare che un utilizzo appropriato dei metadati può ottimizzare l'accessibilità dei dati, rendere più efficiente il recupero delle informazioni e migliorare la gestione del ciclo di vita dei dati stessi.

3.2 Il Dataset: International football results from 1872 to 2023

In questo capitolo ci soffermeremo in maniera particolare su un dataset, trovato sul sito Kaggle. Quest'ultimo consiste in una piattaforma di competizioni di data science e una comunità online di data scientist e professionisti del machine learning, gestita da Google LLC. Su Kaggle gli utenti possono ricercare e pubblicare dataset, sviluppare e testare modelli in un ambiente di data science basato sul web; possono, inoltre, collaborare con altri data scientist e ingegneri del machine learning e partecipare a competizioni volte alla risoluzione di complessi problemi di data science.

Il dataset di nostro interesse include 45.315 risultati di partite di calcio internazionali a partire dalla primissima partita ufficiale nel 1872 fino al 2023. Le partite variano dalla Fifa World Cup alla Copa America fino alle amichevoli regolari. Le partite sono esclusivamente di squadre nazionali maschili; i dati non includono i giochi olimpici, le partite in cui almeno una delle squadre era la squadra B della nazione e gli Under 23. Questo dataset è formato da 3 tabelle; nelle prossime sezioni le esamineremo dettagliatamente. Nella figura 3.1 illustriamo una parte di una tabella (`goalscorers`) che fa parte del dataset preso da Kaggle.

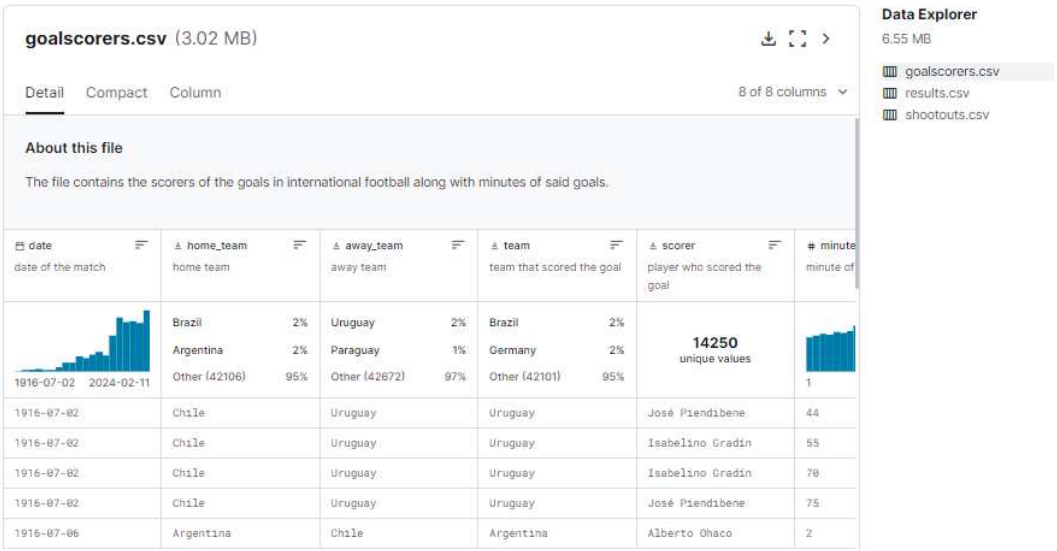


Figura 3.1: Schermata di Kaggle contenente il dataset e le tabelle

3.2.1 Tabella `goalscorers`

La tabella `goalscorers` contiene le informazioni relative ai marcatori delle squadre nazionali nel periodo compreso tra il 1916 e il 2023. Essa è strutturata in 8 colonne, ognuna delle quali rappresenta una specifica tipologia di informazione. La tabella comprende più

di 43000 righe ciascuna delle quali contiene i dati relativi a uno specifico marcatore. Gli attributi sono:

- `date`: questo attributo indica la data della partita. Per esempio giovedì 6 Luglio 1916.
- `home_team`: questo attributo si riferisce al nome della squadra di casa. Un esempio è Brazil.
- `away_team`: questo attributo si riferisce al nome della squadra in trasferta. Un esempio è Chile.
- `team`: questo attributo indica il nome della squadra che ha segnato il gol. Per esempio Brazil.
- `scorer`: questo attributo indica il nome del giocatore che ha segnato il gol. Per esempio Hernando Salazar.
- `minute`: questo attributo indica il numero del minuto in cui ha segnato il marcatore. Per esempio 85.
- `own_goal`: questo attributo è una colonna di tipo booleano e indica se il gol è stato un autogol o no. Per esempio false.
- `penalty`: questo attributo è una colonna di tipo booleano e indica se il gol è stato un rigore o no. Per esempio true.

La Figura 3.2 mostra alcune righe della tabella `goalscorers`.

date	home_team	away_team	team	scorer	minute	own_goal	penalty
domenica 2 luglio 1916	Chile	Uruguay	Uruguay	José Piendibene	44	False	False
domenica 2 luglio 1916	Chile	Uruguay	Uruguay	José Piendibene	75	False	False
domenica 2 luglio 1916	Chile	Uruguay	Uruguay	Isabelino Gradín	70	False	False
domenica 2 luglio 1916	Chile	Uruguay	Uruguay	Isabelino Gradín	55	False	False
giovedì 6 luglio 1916	Argentina	Chile	Argentina	Juan Domingo Brown	60	False	True
giovedì 6 luglio 1916	Argentina	Chile	Argentina	Alberto Marcovecchio	67	False	False
giovedì 6 luglio 1916	Argentina	Chile	Chile	Telésforo Báez	44	False	False
giovedì 6 luglio 1916	Argentina	Chile	Argentina	Alberto Marcovecchio	81	False	False
giovedì 6 luglio 1916	Argentina	Chile	Argentina	Alberto Ohaco	75	False	False
giovedì 6 luglio 1916	Argentina	Chile	Argentina	Alberto Ohaco	2	False	False
giovedì 6 luglio 1916	Argentina	Chile	Argentina	Juan Domingo Brown	62	False	True
sabato 8 luglio 1916	Brazil	Chile	Brazil	Demóstenes Correia de Syllos	29	False	False
sabato 8 luglio 1916	Brazil	Chile	Chile	Hernando Salazar	85	False	False
lunedì 10 luglio 1916	Argentina	Brazil	Brazil	Manoel Alencar Monte	23	False	False
lunedì 10 luglio 1916	Argentina	Brazil	Argentina	José Durand Laguna	10	False	False
mercoledì 12 luglio 1916	Brazil	Uruguay	Brazil	Arthur Friedenreich	8	False	False
mercoledì 12 luglio 1916	Brazil	Uruguay	Uruguay	Isabelino Gradín	58	False	False
mercoledì 12 luglio 1916	Brazil	Uruguay	Uruguay	Jose Tognola	77	False	False
domenica 30 settembre 1917	Uruguay	Chile	Uruguay	Ángel Romano	44	False	False
domenica 30 settembre 1917	Uruguay	Chile	Uruguay	Carlos Scarone	62	False	True
domenica 30 settembre 1917	Uruguay	Chile	Uruguay	Ángel Romano	75	False	False
domenica 30 settembre 1917	Uruguay	Chile	Uruguay	Carlos Scarone	20	False	False

Figura 3.2: Schermata contenente alcune righe della tabella `goalscorers`

3.2.2 Tabella `results`

La tabella `results` contiene tutte le informazioni relative ai risultati delle partite tra le squadre nazionali, nel periodo compreso tra il 1873 e il 2023. Essa è strutturata in 9 colonne, ciascuna delle quali rappresenta un tipo diverso di informazione. La tabella comprende più di 45000 righe, ognuna contenente i dati relativi ad una singola partita. Gli attributi sono:

- `date`: questo attributo indica la data della partita. Per esempio sabato 6 marzo 1875.
- `home_team`: questo attributo si riferisce al nome della squadra di casa. Un esempio è England.
- `away_team`: questo attributo si riferisce al nome della squadra in trasferta. Un esempio è Scotland.
- `home_score`: questo attributo indica il punteggio finale della squadra di casa, inclusi i tempi supplementari, esclusi i rigori. Un esempio è 2.
- `away_score`: questo attributo indica il punteggio finale della squadra ospite, inclusi i tempi supplementari, esclusi i rigori. Un esempio è 1.
- `tournament`: questo attributo indica il nome del torneo a cui partecipano queste squadre. Un esempio è FIFA World Cup.
- `city`: questo attributo indica il nome della città/comune dove si è giocata la partita. Un esempio è London.
- `country`: questo attributo si riferisce al nome del paese dove si è giocata la partita. Un esempio è United States.
- `neutral`: questo attributo è una colonna di tipo booleano e indica se la partita è stata giocata in un campo neutrale o no. Per esempio true.

La Figura 3.3 mostra alcune righe della tabella `results`.

date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
lunedì 15 marzo 1880	Wales	England	2	3	Friendly	Wrexham	Wales	False
sabato 27 marzo 1880	Scotland	Wales	5	1	Friendly	Glasgow	Scotland	False
sabato 26 febbraio 1881	England	Wales	0	1	Friendly	Blackburn	England	False
sabato 12 marzo 1881	England	Scotland	1	6	Friendly	London	England	False
lunedì 14 marzo 1881	Wales	Scotland	1	5	Friendly	Wrexham	Wales	False
sabato 18 febbraio 1882	Northern Ireland	England	0	13	Friendly	Belfast	Ireland	False
sabato 25 febbraio 1882	Wales	Northern Ireland	7	1	Friendly	Wrexham	Wales	False
sabato 11 marzo 1882	Scotland	England	5	1	Friendly	Glasgow	Scotland	False
lunedì 13 marzo 1882	Wales	England	5	3	Friendly	Wrexham	Wales	False
sabato 25 marzo 1882	Scotland	Wales	5	0	Friendly	Glasgow	Scotland	False
sabato 3 febbraio 1883	England	Wales	5	0	Friendly	London	England	False
sabato 24 febbraio 1883	England	Northern Ireland	7	0	Friendly	Liverpool	England	False
sabato 10 marzo 1883	England	Scotland	2	3	Friendly	Sheffield	England	False
lunedì 12 marzo 1883	Wales	Scotland	0	3	Friendly	Wrexham	Wales	False
sabato 17 marzo 1883	Northern Ireland	Wales	1	1	Friendly	Belfast	Ireland	False
sabato 26 gennaio 1884	Northern Ireland	Scotland	0	5	British Home Cha	Belfast	Ireland	False
sabato 9 febbraio 1884	Wales	Northern Ireland	6	0	British Home Cha	Wrexham	Wales	False
sabato 23 febbraio 1884	Northern Ireland	England	1	8	British Home Cha	Belfast	Ireland	False
sabato 15 marzo 1884	Scotland	England	1	0	British Home Cha	Glasgow	Scotland	False
lunedì 17 marzo 1884	Wales	England	0	4	British Home Cha	Wrexham	Wales	False
sabato 29 marzo 1884	Scotland	Wales	4	1	British Home Cha	Glasgow	Scotland	False
sabato 28 febbraio 1885	England	Northern Ireland	4	0	British Home Cha	Manchester	England	False

Figura 3.3: Schermata contenente alcune righe della tabella `results`

3.2.3 Tabella `shootouts`

La tabella `shootouts` contiene tutte le informazioni relative ai rigori delle partite tra le squadre nazionali, nel periodo compreso tra il 1967 e il 2023. Essa è strutturata in 5 colonne, ciascuna delle quali rappresenta un tipo diverso di informazione. La tabella comprende più di 500 righe, ciascuna contenente i dati relativi ad una singola partita. Gli attributi di questa tabella sono:

- `date`: questo attributo indica la data della partita. Per esempio domenica 7 Maggio 1972.
- `home_team`: questo attributo si riferisce al nome della squadra di casa. Un esempio è Senegal.
- `away_team`: questo attributo si riferisce al nome della squadra in trasferta. Un esempio è Ghana.
- `winner`: questo attributo indica la squadra vincitrice dei rigori. Un esempio è Germany.
- `first_shooter`: questo attributo si riferisce alla squadra che batte per prima i rigori. Un esempio è Italy.

La Figura 3.4 mostra alcune righe della tabella `shootouts`.

date	home_team	away_team	winner	first_shooter
martedì 22 agosto 1967	India	Taiwan	Taiwan	
domenica 14 novembre 1971	South Korea	Vietnam Republic	South Korea	
domenica 7 maggio 1972	South Korea	Iraq	Iraq	
mercoledì 17 maggio 1972	Thailand	South Korea	South Korea	
venerdì 19 maggio 1972	Thailand	Cambodia	Thailand	
sabato 21 aprile 1973	Senegal	Ghana	Ghana	
giovedì 14 giugno 1973	Guinea	Mali	Guinea	
giovedì 14 giugno 1973	Mauritius	Tanzania	Mauritius	
giovedì 26 luglio 1973	Malaysia	Kuwait	Malaysia	
giovedì 26 luglio 1973	Cambodia	Singapore	Singapore	
venerdì 27 luglio 1973	Bangladesh	Thailand	Thailand	
sabato 28 luglio 1973	Myanmar	South Korea	Myanmar	
giovedì 9 agosto 1973	India	Vietnam Republic	Vietnam Republic	
giovedì 23 agosto 1973	Algeria	Syria	Syria	
sabato 25 agosto 1973	Algeria	Iraq	Algeria	
giovedì 28 marzo 1974	Qatar	United Arab Emirates	Qatar	
mercoledì 24 luglio 1974	Hong Kong	Indonesia	Indonesia	
giovedì 1 agosto 1974	India	Indonesia	Indonesia	
mercoledì 9 ottobre 1974	Syria	Morocco	Morocco	
venerdì 22 novembre 1974	Libya	Tunisia	Tunisia	
mercoledì 18 dicembre 1974	South Korea	Malaysia	South Korea	
domenica 13 luglio 1975	Morocco	Ghana	Morocco	
domenica 9 novembre 1975	Kenya	Malawi	Kenya	

Figura 3.4: Schermata contenente alcune righe della tabella `shootouts`

3.3 Attività di ETL

In questa sezione esamineremo approfonditamente le operazioni di ETL (Figura 3.5) eseguite sul dataset relativo ai risultati delle partite di calcio delle squadre nazionali. L'attività di ETL (Extract, Transform, Load) rappresenta uno dei passaggi preliminari, fondamentali nell'ambito dell'analisi dei dati. Questo processo comprende un insieme di funzioni che

estraggono, trasformano e caricano i dati da una o più sorgenti verso un sistema di destinazione. L'obiettivo principale dell'ETL è convertire i dati grezzi in informazioni operative utili per le attività di Business Intelligence. L'acronimo "ETL" si riferisce alle tre fasi del processo, descritte di seguito:

- *Extract*: durante questa fase, i dati vengono raccolti da diverse fonti, come database, file e fogli di calcolo, e preparati per la successiva elaborazione.
- *Transform*: in questa fase, i dati raccolti vengono convertiti in un formato compatibile e ottimizzato per il sistema di destinazione.
- *Load*: la fase di caricamento prevede l'importazione dei dati trasformati nel sistema di destinazione, che solitamente è un Data Warehouse, un Data Mart o un software di Business Intelligence.

Le operazioni di ETL sono cruciali per la Business Intelligence, la gestione, la migrazione e l'integrazione dei dati, oltre che per la preparazione degli stessi per l'analisi. Esse permettono alle aziende di unificare dati provenienti da diversi sistemi, trasformandoli in informazioni utili per prendere decisioni informate. Nelle sezioni seguenti, analizzeremo queste tre fasi dettagliatamente.

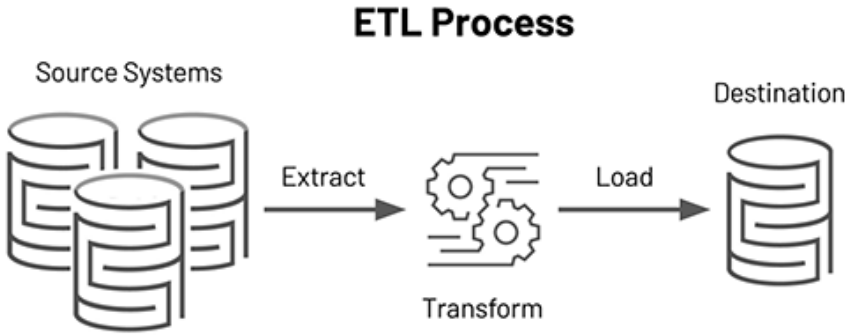


Figura 3.5: Attività di ETL

3.3.1 Extract

Nel processo di estrazione dei dati, il primo passaggio consiste nell'identificazione del dataset che sarà oggetto di analisi. Generalmente, si tende a selezionare una varietà di fonti di dati per individuare schemi e relazioni tra di essi. È importante sottolineare che le raccolte di dati possono provenire sia dall'interno dell'organizzazione, come i database aziendali, che dall'esterno, come i social media. Una volta individuati i set di dati pertinenti, si procede al loro recupero e alla loro memorizzazione; ciò garantisce la disponibilità di una copia di tutti i dati necessari per l'analisi, che potranno essere successivamente modellati in modo ottimale. L'estrazione dei dati può essere eseguita in vari modi, tra cui lo sviluppo di programmi su misura, l'uso di uno dei numerosi strumenti di ETL presenti sul mercato, o una combinazione di entrambi. Alcuni strumenti di ETL moderni possono semplificare notevolmente le operazioni legate a questa fase, eliminando la necessità di scrivere righe di codice, poiché tali attività sono gestite automaticamente da questi strumenti. Ciò può ridurre significativamente la quantità di personale richiesto per questo passaggio. Tuttavia, esiste un compromesso in termini di costi e benefici, poiché questi strumenti tendono ad essere più

costosi, a fronte della disponibilità gratuita di linguaggi di programmazione come Python e Java.

3.3.2 Transform

Durante questa fase vengono applicate diverse funzioni o metodi per manipolare i dati estratti. Questa operazione è fondamentale per pulire i dati, combinare dati provenienti da diverse fonti, suddividere informazioni in più tabelle e altre operazioni simili. L'obbiettivo è trasformare i dati in un formato facilmente utilizzabile e che soddisfi efficacemente le esigenze aziendali. Durante questa fase, il trattamento dei dati grezzi può comprendere le seguenti operazioni:

- validazione e pulizia dei dati per correggere errori;
- mappatura dei dati nel formato richiesto;
- eliminazione dei duplicati;
- filtraggio dei dati superflui;
- verifica dell'uniformità dei formati per garantire la coerenza.

Inoltre, possono essere applicati step avanzati di trasformazione dei dati in base alle necessità. Esempi di tali step sono:

- la suddivisione dei dati in più colonne;
- la fusione di tabelle;
- la condensazione dei dati per ridurre le dimensioni del dataset.

Attraverso questi step, ciò che inizialmente era un insieme di materiale inutilizzabile viene convertito in un prodotto di dati, pronto per la fase finale del processo di ETL, ossia il caricamento.

3.3.3 Load

Una volta concluse le fasi di estrazione e trasformazione dei dati, l'operazione finale consiste nel caricare gli stessi, trasformati nel Data Warehouse. Questo processo è preciso, continuo e automatizzato; di norma, il caricamento dei dati avviene in blocchi. Le possibili modalità di caricamento sono le seguenti:

- *Caricamento totale dei dati*: questo tipo di caricamento avviene solitamente nella fase iniziale; coinvolge l'estrazione e la trasformazione dell'intero set di dati dalla fonte al data warehouse.
- *Caricamento dei dati a blocchi*: quando il set di dati è molto grande, questi ultimi vengono caricati in blocchi e a intervalli periodici.
- *Caricamento incrementale dei dati*: questo processo comporta il caricamento periodico solo dei dati aggiornati tra il sistema di origine e il sistema di destinazione.
- *Caricamento incrementale in streaming*: questo metodo prevede lo streaming continuo dei dati, ma è adatto solo per set di dati di dimensioni più contenute.

3.4 ETL sul Dataset: International football results from 1872 to 2023

In questa sezione illustriamo in modo dettagliato le procedure di Estrazione, Trasformazione e Caricamento applicate al dataset relativo ai risultati delle partite di calcio delle squadre nazionali. Offriamo un'analisi approfondita delle tre fasi principali, corredata da numerosi screenshot per descrivere i vari passaggi, le interfacce utilizzate e le azioni eseguite. Particolare attenzione è stata riservata alle operazioni di pulizia e trasformazione dei dati, effettuate tramite l'editor di query di Power BI, noto come Power Query.

3.4.1 Extract

Come precedentemente detto, l'estrazione costituisce la fase iniziale del processo di ETL. Per questa attività utilizziamo Power BI, che ci offre un'ampia gamma di opzioni. L'interfaccia principale del software, illustrata nella Figura 3.6, presenta quattro delle opzioni più comunemente utilizzate, oltre alla possibilità di effettuare ulteriori selezioni. Ciò si può vedere chiaramente nella Figura 3.7, situata nella parte superiore dell'interfaccia di Power BI. Attraverso entrambe le opzioni, accediamo alla schermata di raccolta dei dati, come mostrato nella Figura 3.8. In questa finestra, possiamo selezionare la nostra fonte di interesse da una varietà di risorse compatibili con Power BI. Una volta effettuata la selezione, procediamo a connettere la fonte scelta a Power BI utilizzando la funzione "Connetti".

Figura 3.6: Schermata iniziale di Power BI per la selezione dei dati

Figura 3.7: Barra superiore di Power BI dedicata alla raccolta di dati provenienti da diverse fonti

Una volta stabilita la connessione, si aprirà la schermata dello strumento di navigazione (Figura 3.9), dove è possibile visualizzare in anteprima i dati, eseguire un'analisi preliminare e decidere se trasformare o caricare direttamente i dati, qualora quest'ultimi siano già pronti per l'analisi. Nel nostro caso, scegliamo di trasformare i dati prima del caricamento; pertanto, inizieremo con Power Query per tutte e tre le tabelle.

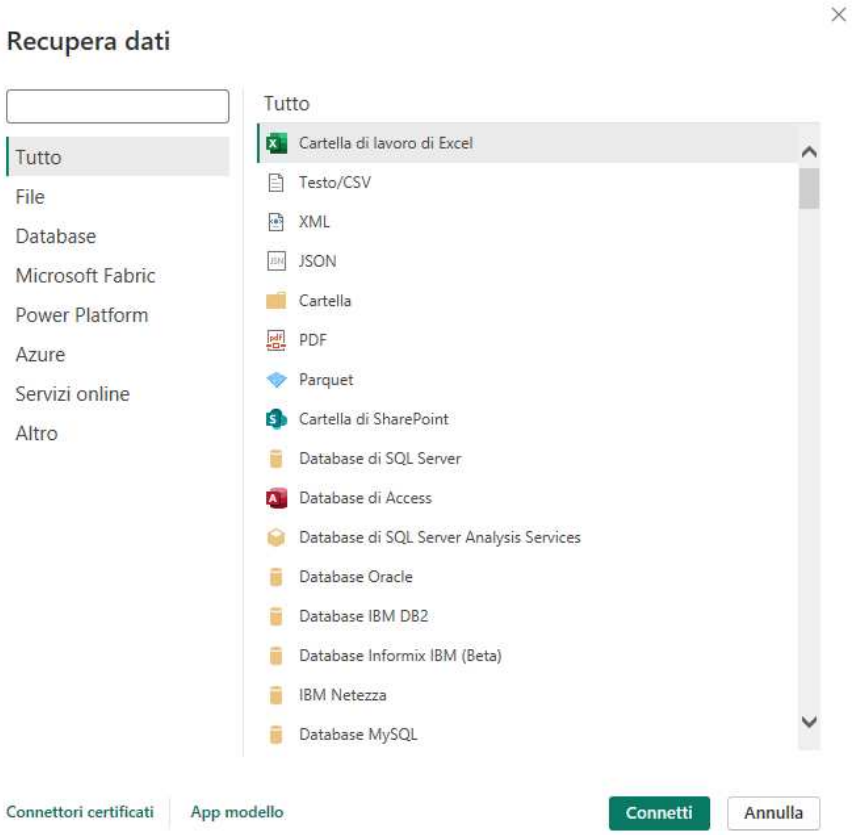
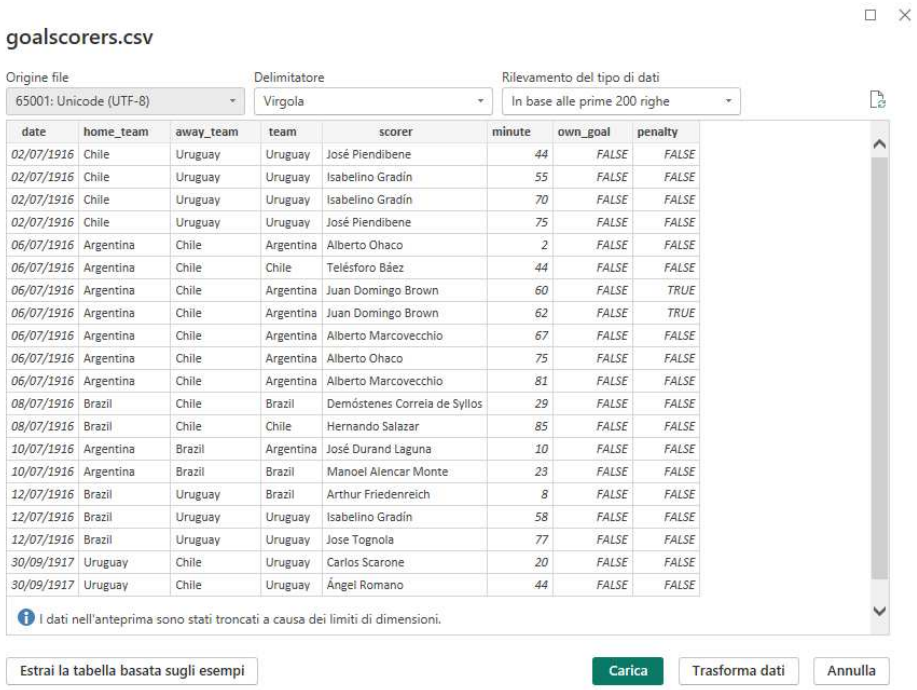


Figura 3.8: Finestra "Recupera dati" in Power BI Desktop



date	home_team	away_team	team	scorer	minute	own_goal	penalty
02/07/1916	Chile	Uruguay	Uruguay	José Piendibene	44	FALSE	FALSE
02/07/1916	Chile	Uruguay	Uruguay	Isabelino Gradín	55	FALSE	FALSE
02/07/1916	Chile	Uruguay	Uruguay	Isabelino Gradín	70	FALSE	FALSE
02/07/1916	Chile	Uruguay	Uruguay	José Piendibene	75	FALSE	FALSE
06/07/1916	Argentina	Chile	Argentina	Alberto Ohaco	2	FALSE	FALSE
06/07/1916	Argentina	Chile	Chile	Telésforo Báez	44	FALSE	FALSE
06/07/1916	Argentina	Chile	Argentina	Juan Domingo Brown	60	FALSE	TRUE
06/07/1916	Argentina	Chile	Argentina	Juan Domingo Brown	62	FALSE	TRUE
06/07/1916	Argentina	Chile	Argentina	Alberto Marcovecchio	67	FALSE	FALSE
06/07/1916	Argentina	Chile	Argentina	Alberto Ohaco	75	FALSE	FALSE
06/07/1916	Argentina	Chile	Argentina	Alberto Marcovecchio	81	FALSE	FALSE
08/07/1916	Brazil	Chile	Brazil	Demóstenes Correia de Syllos	29	FALSE	FALSE
08/07/1916	Brazil	Chile	Chile	Hernando Salazar	85	FALSE	FALSE
10/07/1916	Argentina	Brazil	Argentina	José Durand Laguna	10	FALSE	FALSE
10/07/1916	Argentina	Brazil	Brazil	Manoel Alencar Monte	23	FALSE	FALSE
12/07/1916	Brazil	Uruguay	Brazil	Arthur Friedenreich	8	FALSE	FALSE
12/07/1916	Brazil	Uruguay	Uruguay	Isabelino Gradín	58	FALSE	FALSE
12/07/1916	Brazil	Uruguay	Uruguay	Jose Tognola	77	FALSE	FALSE
30/09/1917	Uruguay	Chile	Uruguay	Carlos Scarone	20	FALSE	FALSE
30/09/1917	Uruguay	Chile	Uruguay	Ángel Romano	44	FALSE	FALSE

Figura 3.9: Finestra "strumento di navigazione" in Power BI Desktop

3.4.2 Transform

Completato il processo di estrazione, si procede con la fase di trasformazione. In questo stadio, i dati vengono principalmente puliti e strutturati in una forma che consente la generazione di report significativi. In primo luogo, si verifica che tutti gli attributi, ovvero le colonne, siano espressi nel tipo corretto. Nel nostro caso specifico, è stato necessario effettuare una conversione di tipo per alcuni attributi. Per ogni tabella, il tipo del campo `date` è stato convertito in `datetime`; oltre a questo, non sono stati effettuati altri interventi. Successivamente, verrà esaminata l'accuratezza delle intestazioni; sotto questo aspetto, non sono emersi problemi. A questo punto, si potranno analizzare i campi singolarmente. Nella tabella `results`, come vediamo nella Figura 3.10, il campo `date` è formato dal giorno, dal mese e dall'anno della partita. L'anno è un'informazione rilevante perchè ci servirà per i nostri report; per questo motivo, con una formula DAX, verrà estratto l'anno dalla colonna date. Di seguito, nella Figura 3.11, possiamo osservare la nuova colonna `year`.

date
domenica 12 febbraio 1905
domenica 4 aprile 1909
domenica 1 dicembre 1912
domenica 12 gennaio 1913
giovedì 1 maggio 1913
giovedì 13 maggio 1915
mercoledì 12 luglio 1916
giovedì 29 maggio 1919
domenica 19 settembre 1920
domenica 24 ottobre 1920
lunedì 4 aprile 1921
domenica 18 dicembre 1921

Figura 3.10: Schermata di una porzione della colonna `date`

Year
1905
1909
1912
1913
1913
1915
1916
1919
1920

Figura 3.11: Finestra di una porzione della colonna `year`

3.4.3 Load

Dopo aver effettuato le modifiche richieste tramite il processo di trasformazione, si procede alla fase di caricamento dei dati. In questa circostanza, i dati vengono trasferiti dall'editor Power Query direttamente in Power BI, dove saranno disponibili per l'elaborazione dei report. Per finalizzare questa operazione, è sufficiente selezionare l'opzione "Chiudi e applica", come mostrato nella Figura 3.12.

Figura 3.12: Opzione "Chiudi e applica" di Power Query

CAPITOLO 4

Analisi generali

In questo capitolo analizzeremo, in modo generale, le procedure di analisi dei dati eseguite sul Dataset riguardanti i risultati delle squadre nazionali.

4.1 Panoramica generale del Dataset

In questa sezione ci occuperemo di descrivere le analisi generali di Data Analytics effettuate sul Dataset riguardanti i risultati delle squadre nazionali. In particolare, come possiamo vedere nella Figura 4.1, analizzeremo dei report come grafici a barre, filtri, ma anche schede e indicatori KPI. Per quanto riguarda le schede, faremo delle considerazioni generali su ognuna di esse. Gli indicatori che abbiamo preso in considerazione sono i seguenti:

- `totale partite`: questo valore rappresenta il numero totale di partite giocate memorizzate nel Dataset. È una misura del volume complessivo di dati relativi agli incontri di calcio nazionali raccolti nel periodo 1872-2023.
- `totale goal`: indica il numero totale di goal segnati in tutte le partite presenti nel Dataset. Questo dato fornisce una panoramica dell'efficacia offensiva complessiva nelle partite considerate.
- `media goal per squadra di casa`: questo valore medio riflette il numero di goal segnati per partita dalla squadra che giocava in casa. Un valore medio di 1.56 suggerisce una tendenza delle squadre di casa a segnare più goal rispetto a quelle in trasferta.
- `media goal per squadra in trasferta`: questo valore rappresenta il numero di goal segnati per partita dalla squadra in trasferta. Un valore di 1.19 dimostra che le squadre in trasferta tendono a segnare di meno rispetto a quelle che giocano in casa.
- `goal per anno`: il grafico associato a questo indicatore evidenzia la distribuzione dei goal segnati per anno. Le oscillazioni nel numero di goal per anno possono riflettere cambiamenti nelle regole del gioco, negli stili di gioco o altri fattori esterni come la frequenza delle partite internazionali.

La dashboard include anche dei filtri per anno, città, paese e torneo, che permettono agli utenti di esplorare i dati in modo più dettagliato e mirato.

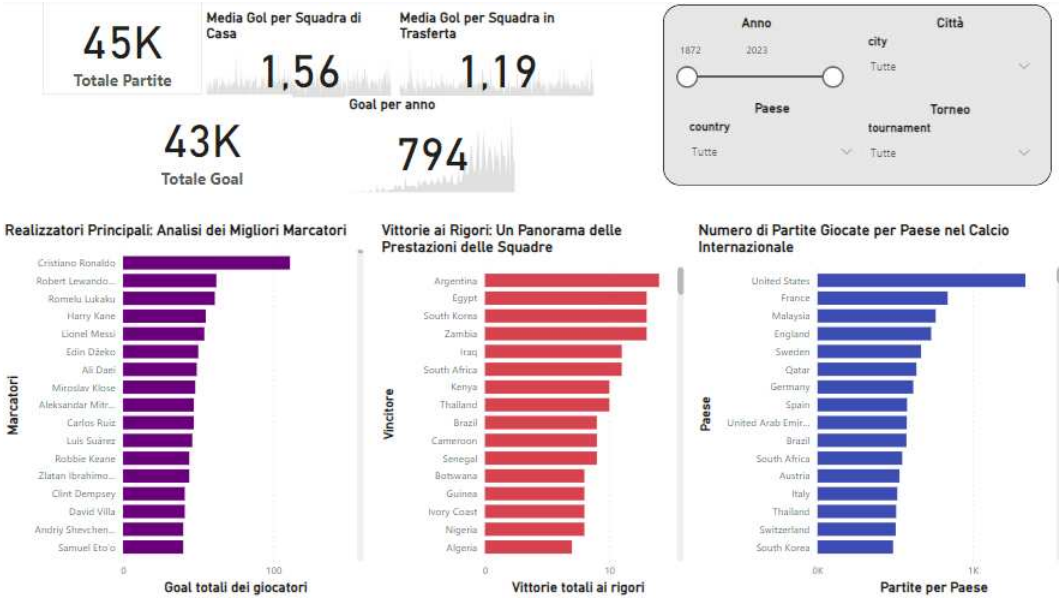


Figura 4.1: Panoramica generale del dataset

4.1.1 Report sui principali marcatori

Nella Figura 4.2 abbiamo un grafico a barre che elenca i giocatori con il maggior numero di goal segnati. Cristiano Ronaldo è il miglior marcatore, un dato che conferma la sua fama e la sua straordinaria capacità di segnare nel corso della sua carriera internazionale. Robert Lewandowski e Romelu Lukaku seguono nella lista, entrambi noti per le loro abilità offensive e la costanza nel segnare. La presenza di giocatori come Lionel Messi, Harry Kane e Zlatan Ibrahimovic evidenzia come i migliori marcatori provengano dalle principali leghe europee, suggerendo un alto livello di competizione e di esposizione mediatica. Questi dati possono essere utilizzati per analizzare l'efficacia delle diverse strategie offensive adottate dai vari paesi, e come queste ultime influenzano le prestazioni individuali dei giocatori. Il grafico può anche suggerire la necessità di sviluppare talenti a livello giovanile; infatti, molti dei migliori marcatori hanno iniziato le loro carriere internazionali in giovane età.

4.1.2 Report sulle vittorie ai rigori

Nella Figura 4.3 troviamo un grafico a barre che indica il numero totale di vittorie ai rigori per ogni squadra. Argentina ed Egitto emergono come squadre con il maggior numero di vittorie ai rigori, suggerendo una forte preparazione psicologica e tecnica in situazioni di alta pressione. Sud Corea, Zambia e Iraq seguono, mostrando che anche nazioni meno conosciute nel panorama calcistico internazionale possono eccellere in situazioni specifiche come i calci di rigore. Le squadre africane e asiatiche hanno una rappresentazione significativa in questo grafico, indicando una possibile specializzazione o strategia mirata in questi paesi. Il numero di vittorie ai rigori può riflettere non solo l'abilità tecnica dei giocatori, ma anche l'efficacia dell'allenamento e della preparazione mentale. Questi dati possono essere utili per le squadre e per gli allenatori che cercano di migliorare le loro prestazioni ai rigori perchè indicano quali squadre hanno strategie vincenti che potrebbero essere studiate o emulate. Il grafico, inoltre, può suggerire l'importanza della gestione dello stress nei momenti cruciali di una partita.

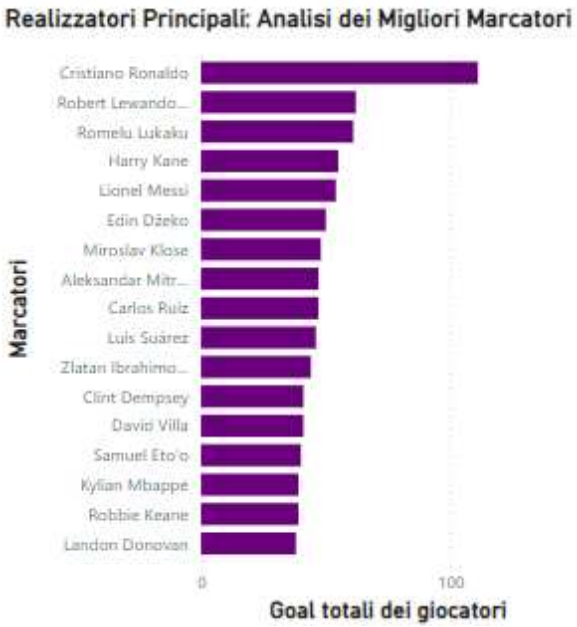


Figura 4.2: Grafico a barre dei principali marcatori

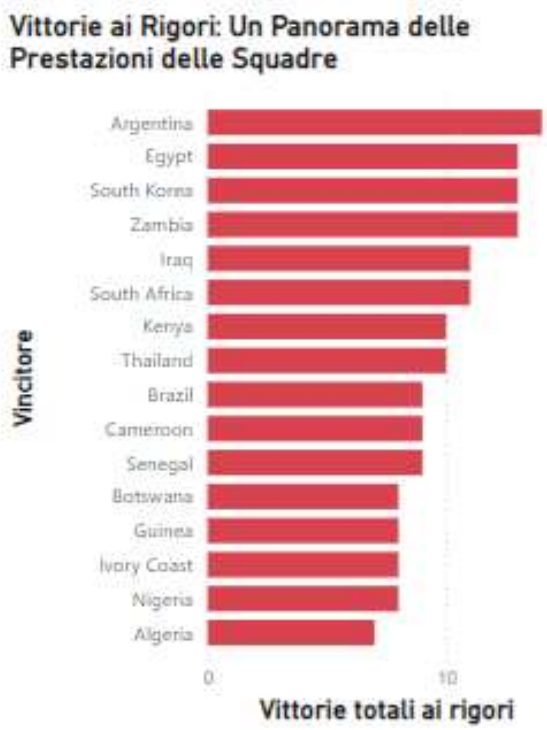


Figura 4.3: Grafico a barre delle vittorie ai rigori

4.1.3 Report sulle partite giocate da ogni paese

Nella Figura 4.4 analizzeremo un ulteriore grafico a barre che mostra il numero di partite internazionali giocate da ciascun paese. Gli Stati Uniti sono in cima alla lista con il maggior numero di partite giocate; ciò indica una forte presenza nel calcio internazionale, che può essere attribuita alla crescita di questo sport negli Stati Uniti negli ultimi decenni. Seguono Francia e Malesia, con la Francia che è una potenza storica del calcio e la Malesia che eviden-

zia una partecipazione attiva nelle competizioni internazionali. La presenza di paesi come Inghilterra, Svezia e Germania sottolinea la lunga tradizione e la continua partecipazione di queste nazioni in tornei internazionali. La rappresentazione di paesi di diversi continenti (ad esempio, Qatar, Brasile, Sudafrica, Thailandia) mostra la globalizzazione del calcio e come esso si sia diffuso in tutto il mondo. Un alto numero di partite giocate può indicare non solo la competitività e l'esperienza della squadra nazionale, ma anche l'investimento del paese nel calcio come sport. Questo grafico può essere utile come indicatore della popolarità e dello sviluppo del calcio nei vari paesi, influenzando politiche sportive e investimenti. Esso può anche suggerire opportunità per analizzare le prestazioni a lungo termine e l'evoluzione delle squadre nazionali, individuando tendenze e cambiamenti nel tempo.

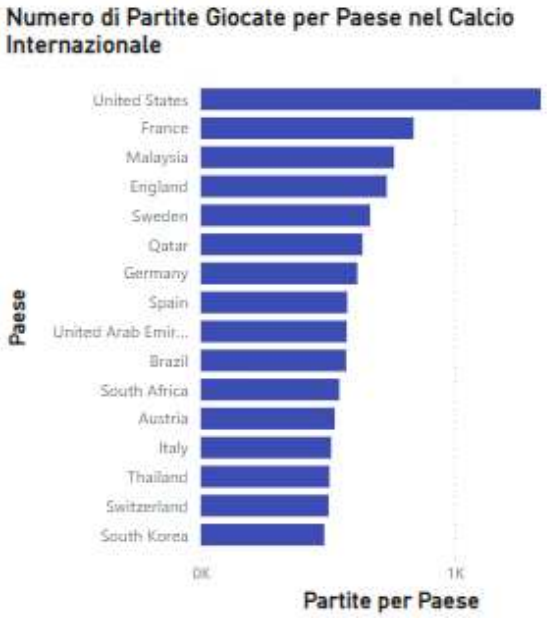


Figura 4.4: Grafico a barre delle partite giocate da ogni paese

CAPITOLO 5

Analisi specifiche

In questo capitolo analizzeremo, in modo specifico, le procedure di analisi dei dati eseguite sul Dataset riguardanti i risultati delle squadre nazionali. Ci soffermeremo, in particolare, su tre dashboard: la prima riguardante la tabella results, la seconda relativa alla tabella goalscorers e la terza attinente alle 5 nazionali principali dal 2000 al 2023.

5.1 Analisi specifiche tabella `results`

In questa sezione, effettueremo un'indagine dettagliata sui report inerenti alla tabella `results`, con l'obbiettivo di analizzare i 5 tornei principali, facendo un confronto fra di loro. Come possiamo vedere nella Figura 5.1, abbiamo diversi grafici, come grafici a barre in pila e istogrammi a colonne raggruppate; questi ultimi verranno analizzati nelle prossime sottosezioni.

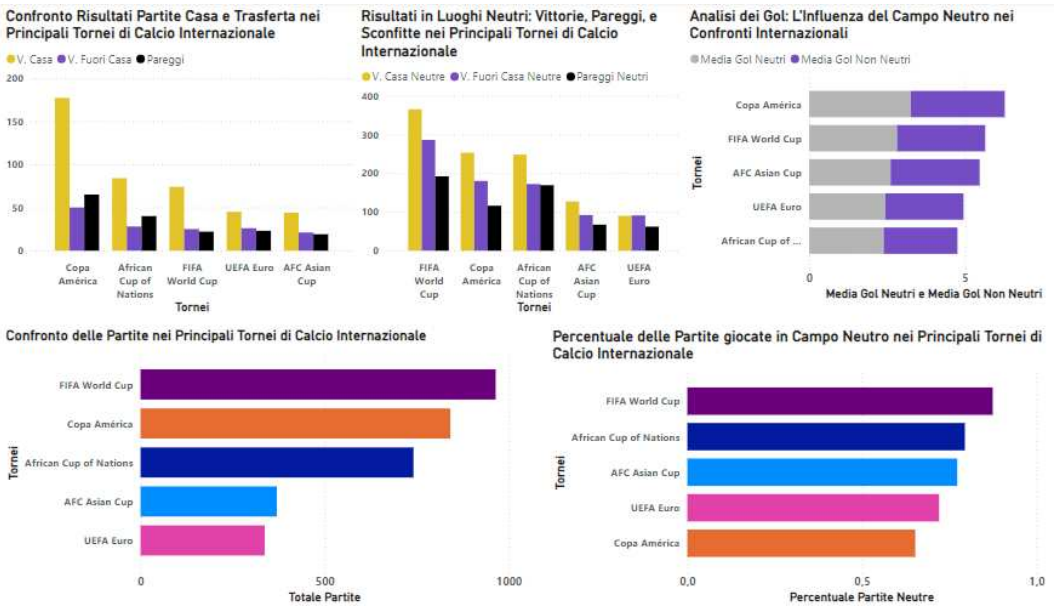


Figura 5.1: Dashboard riguardante la tabella `results`

5.1.1 Report sui risultati delle partite in casa e in trasferta

Nella Figura 5.2 troviamo un istogramma a colonne raggruppate che mostra il confronto tra vittorie in casa, vittorie fuori casa e pareggi nei principali tornei internazionali. Questo grafico evidenzia la predominanza delle vittorie casalinghe, una tendenza che sottolinea il concetto di "vantaggio di casa". Tuttavia, l'analisi rivela anche che la Copa America presenta un numero insolitamente elevato di vittorie esterne, suggerendo che, in questo torneo, le squadre ospiti ottengono un rendimento sorprendentemente alto. Questo fenomeno potrebbe essere esplorato ulteriormente per comprendere se esso sia il risultato di specifiche condizioni del torneo, come il clima, la tattica delle squadre, oppure di altri fattori culturali e psicologici.

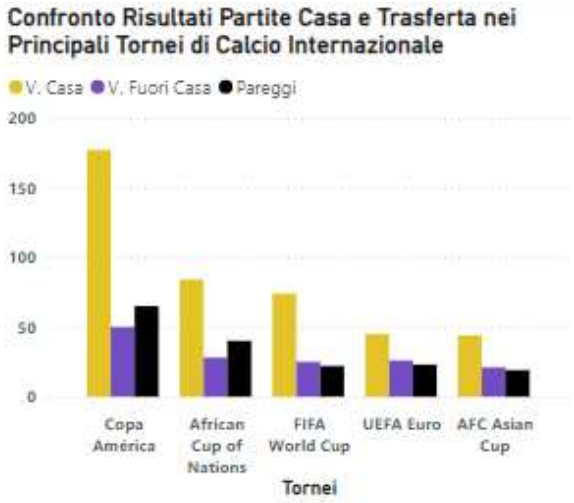


Figura 5.2: Istogramma riguardante i risultati delle partite in casa e trasferta

5.1.2 Report sui risultati delle partite in luoghi neutri

Nella Figura 5.3 consideriamo un grafico che confronta le vittorie in casa, le vittorie fuori casa e i pareggi in luogo neutro nei 5 tornei principali. I risultati in luoghi neutri mostrano una distribuzione più uniforme tra vittorie, pareggi e sconfitte; ciò implica che l'eliminazione del vantaggio di casa genera una competizione più equilibrata e meno prevedibile. Questo può essere un indicatore di come la neutralità del luogo di gioco contribuisce a livellare le differenze tra le squadre, enfatizzando l'importanza della preparazione tecnica e tattica e riducendo l'effetto psicologico del supporto dei tifosi.

5.1.3 Report sulla media dei goal neutri e non neutri

Nella Figura 5.4 analizziamo un grafico a barre impilato che mette in evidenza la media dei goal segnati in partite giocate in campo neutro rispetto a quelle non neutre, per ogni torneo principale. I dati suggeriscono che il vantaggio di giocare in casa potrebbe essere meno influente nel conteggio dei goal rispetto a quanto comunemente percepito, specialmente in tornei come la Copa America e l'African Cup of Nations. Questo potrebbe essere interpretato come un'indicazione della crescente capacità delle squadre di adattarsi a contesti diversi dal proprio ambiente, forse grazie a migliori strategie di preparazione o a una maggiore familiarità con condizioni di gioco variabili.

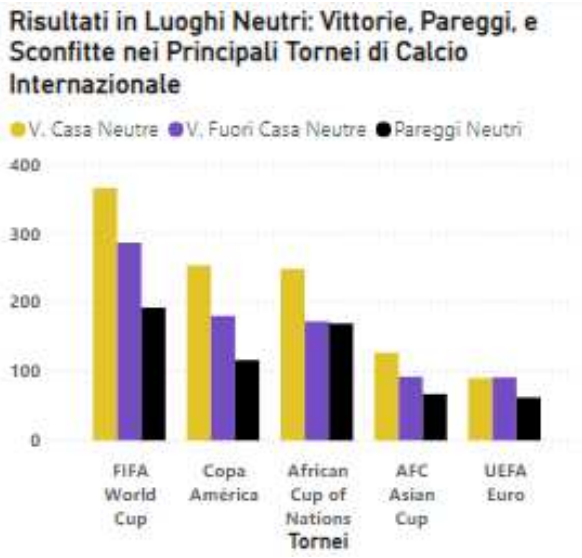


Figura 5.3: Istogramma riguardante i risultati delle partite in luoghi neutri

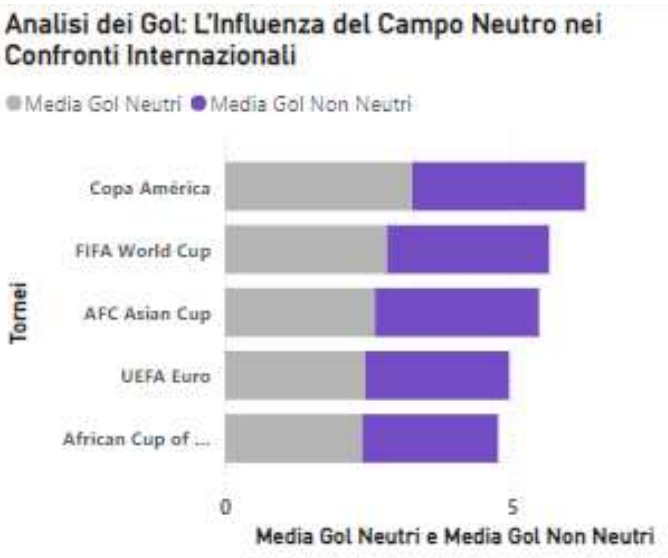


Figura 5.4: Grafico a barre riguardante la media dei goal neutri e non neutri

5.1.4 Report sul confronto delle partite giocate

Nella Figura 5.5 troviamo un grafico a barre che rappresenta il numero totale di partite giocate per ogni torneo principale. In questo report emerge chiaramente che la FIFA World Cup detiene il primato per il numero di incontri disputati. Questo dato è indicativo della rilevanza e della tradizione che il torneo, nato nel 1930, ha nel panorama calcistico mondiale. Inoltre, la Copa America e l'African Cup of Nations seguono con numeri significativi, riflettendo la passione e l'importanza del calcio in questi continenti. Tuttavia, la maggiore frequenza di partite potrebbe anche essere correlata alla struttura dei tornei, che può prevedere fasi preliminari più estese o un maggior numero di squadre partecipanti.

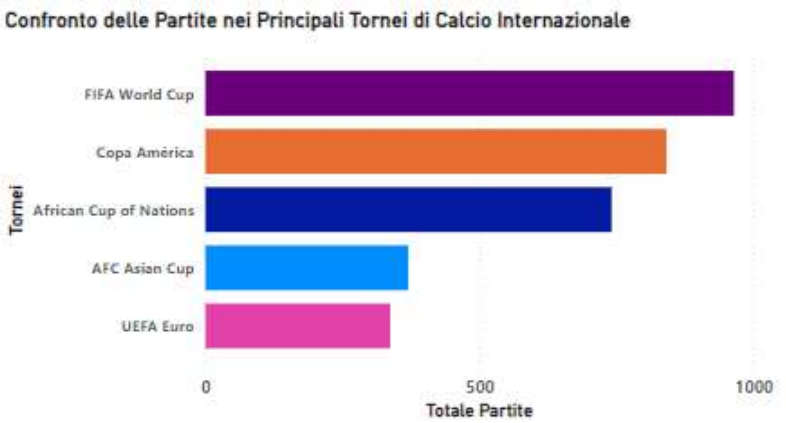


Figura 5.5: Grafico a barre riguardante il numero di partite giocate

5.1.5 Report sulla percentuale delle partite giocate in campo neutro

Nella Figura 5.6 esaminiamo un grafico a barre che mostra la percentuale di partite giocate in campo neutro rispetto al totale delle partite per ogni torneo. In questo report si nota che la Fifa World Cup si distingue notevolmente dagli altri tornei. Questo potrebbe essere attribuito al formato unico del torneo che, nelle edizioni recenti, ha visto vari paesi ospitanti contemporaneamente, eliminando così il classico "vantaggio di casa". Questo formato potrebbe essere visto come un tentativo di equilibrare il campo di gioco, dando a tutte le squadre uguali opportunità, indipendentemente dalla loro base geografica.

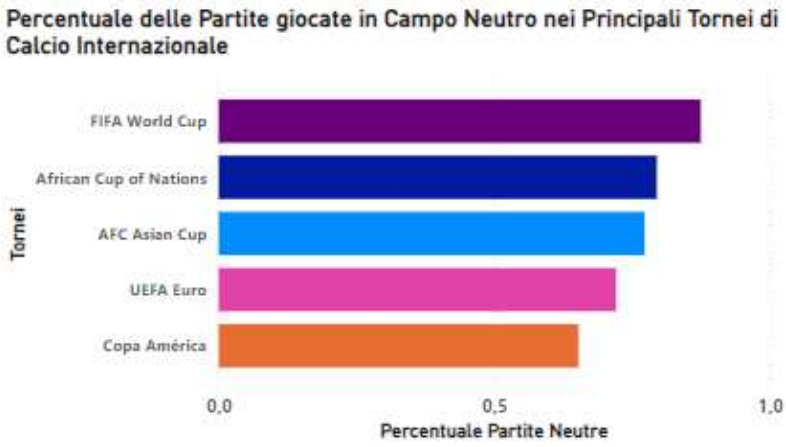


Figura 5.6: Grafico a barre riguardante la percentuale delle partite giocate in campo neutro

5.2 Analisi specifiche tabella `goalscorers`

In questa sezione, effettueremo un'indagine dettagliata sui report riguardanti la tabella `goalscorers`, con l'obbiettivo di analizzare i 5 marcatori principali, facendo un confronto fra di loro. Come possiamo vedere nella Figura 5.7, abbiamo diversi grafici, come grafici a barre in pila, grafici a linee e istogrammi a colonne raggruppate; sono presenti, anche, due filtri, uno per l'anno e uno per il torneo. Questi grafici verranno analizzati in dettaglio nelle prossime sottosezioni.

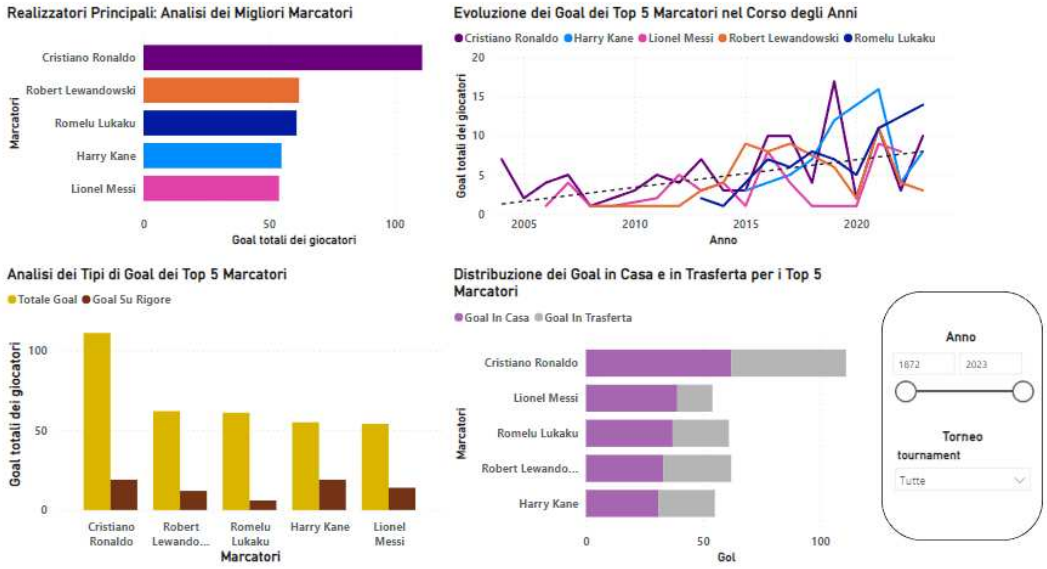


Figura 5.7: Dashboard riguardante la tabella `goalscorers`

5.2.1 Report sui realizzatori principali

Nella Figura 5.8 analizziamo un grafico a barre che mostra il totale dei gol segnati dai cinque migliori marcatori: Cristiano Ronaldo, Robert Lewandowski, Romelu Lukaku, Harry Kane e Lionel Messi. È importante notare la predominanza di Ronaldo, che non solo quantifica la sua eccezionale propensione al goal ma può anche riflettere le sue qualità di resilienza e costanza nel corso di una carriera prolungata ai massimi livelli internazionali. Il grafico, inoltre, suggerisce una competizione più serrata tra gli altri quattro giocatori.

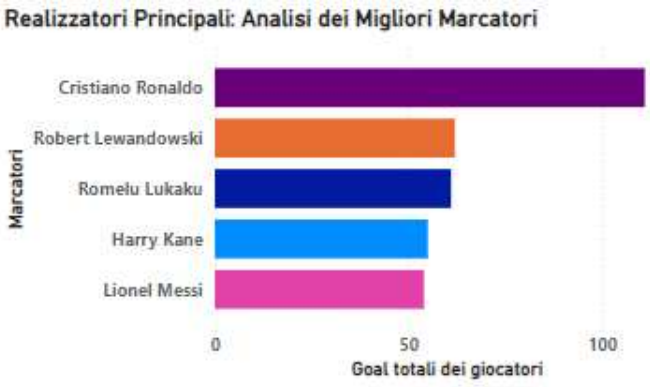


Figura 5.8: Grafico a barre riguardante il totale dei goal segnati

5.2.2 Report sull'evoluzione dei goal nel tempo

Nella Figura 5.9 troviamo un grafico a linee che mostra l'andamento dei goal segnati dai cinque migliori marcatori nel corso degli anni. Esso permette di visualizzare la progressione delle prestazioni dei giocatori nel corso del tempo e indica anche i periodi di picco o di calo della forma. Dal grafico emergono varie tendenze, come il costante apporto di Ronaldo e l'ascesa di Lewandowski negli anni recenti. Questa visualizzazione è cruciale per analizzare la consistenza e la longevità dei giocatori.

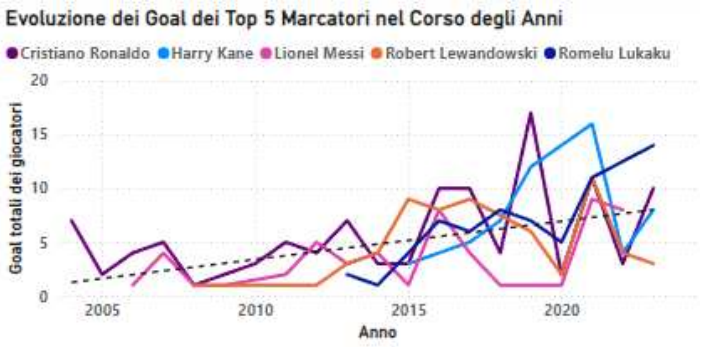


Figura 5.9: Grafico a linee riguardante l'evoluzione dei goal nel tempo per i 5 migliori marcatori

5.2.3 Report sull'analisi dei tipi di goal

Nella Figura 5.10 analizziamo un grafico a barre che suddivide i goal totali in due categorie: goal totali e goal su rigore, per ciascuno dei cinque marcatori. Esso offre una prospettiva sulle capacità realizzative intrinseche dei giocatori al di là delle situazioni di gioco statiche. La quantità di goal su rigore può anche riflettere la fiducia riposta nel giocatore e la sua capacità di gestire la pressione in momenti critici del match.

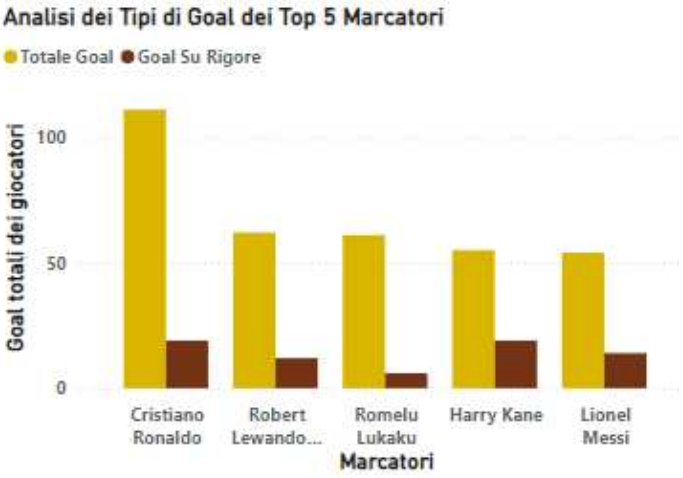


Figura 5.10: Istogramma riguardante i goal totali e i goal su rigore

5.2.4 Report sulla distribuzione dei goal in casa e in trasferta

Nella Figura 5.11 esaminiamo un grafico a barre impilate che offre una comparazione diretta tra i goal segnati in casa e in trasferta dai migliori marcatori; esso dimostra la capacità di ogni giocatore di mantenere l'efficacia realizzativa in ambienti differenti. Un'analisi di questo tipo può essere indicativa dell'abilità di un giocatore di esibirsi sotto diverse pressioni ambientali e di tifo. È particolarmente rilevante osservare giocatori come Ronaldo e Lewandowski che mantengono un buon bilancio tra prestazioni casalinghe e trasferte, suggerendo una grande capacità di adattamento e costanza.

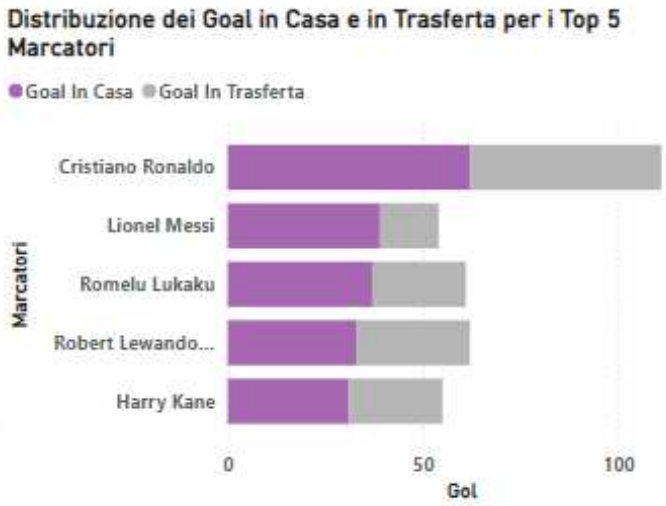


Figura 5.11: Grafico a barre riguardante i goal in casa e in trasferta

5.3 Analisi specifiche sulle nazionali dal 2000 al 2023

In questa sezione, ci soffermeremo sui report riguardanti le 5 nazionali principali, con l'obbiettivo di analizzare queste ultime nel periodo che va dal 2000 al 2023, facendo un confronto fra di loro. Come possiamo vedere nella Figura 5.12, troviamo diversi grafici come: grafici a barre in pila, grafici a linee, istogrammi a colonne raggruppate e grafici a ciambella; è presente, anche, un filtro relativo agli anni. Analizzeremo in dettaglio ciascuno di questi grafici nelle prossime sottosezioni.

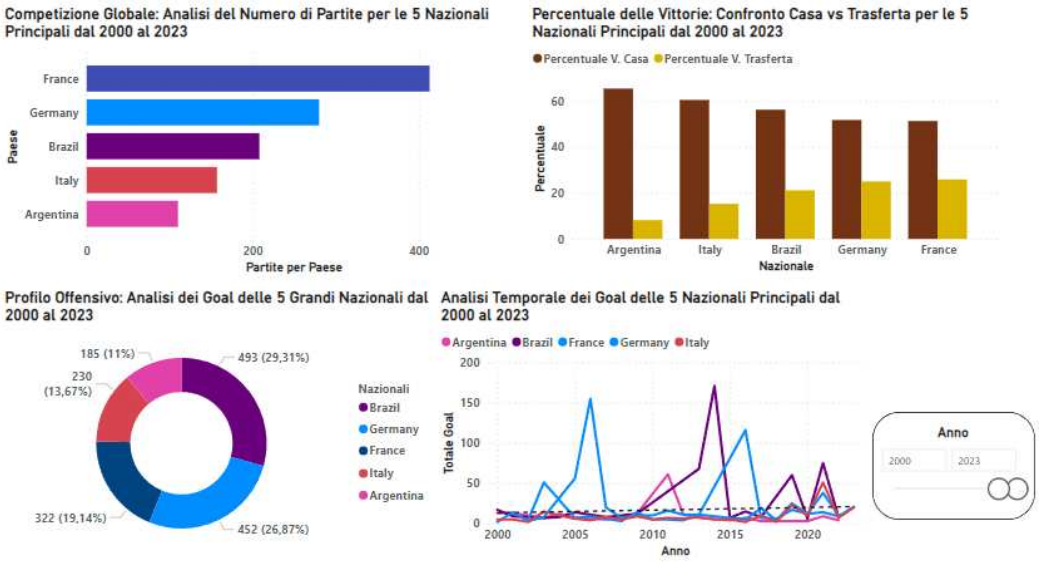


Figura 5.12: Dashboard riguardante le nazionali principali dal 2000 al 2023

5.3.1 Report sul numero di partite

Nella Figura 5.13 analizziamo un grafico a barre che mostra il numero totale di partite giocate dalle cinque nazionali principali (Francia, Germania, Brasile, Italia e Argentina) dal 2000 al 2023. Francia e Germania hanno giocato il maggior numero di partite nel periodo con-

siderato, con una maggiore presenza in tornei internazionali e amichevoli. L'Argentina ha giocato il minor numero di partite tra le cinque nazionali, indicando una possibile differenza nel numero di competizioni. È utile per avere un'idea immediata del volume di esperienza internazionale che ciascuna squadra possiede, indicativo, forse, di una storia calcistica più lunga o di una maggiore frequenza di incontri.

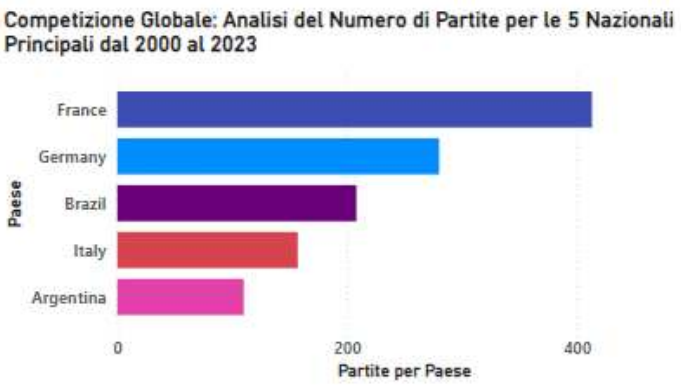


Figura 5.13: Grafico a barre riguardante le partite giocate dalle 5 nazionali principali

5.3.2 Report sulla percentuale delle vittorie in casa e in trasferta

Nella Figura 5.14 troviamo un grafico a colonne affiancate che permette un confronto diretto delle percentuali di vittoria in casa e in trasferta per le cinque squadre. Tutte le nazionali mostrano una tendenza a vincere di più in casa rispetto alle trasferte. Argentina e Italia hanno una differenza significativa tra le percentuali di vittorie in casa e in trasferta, indicando una maggiore dipendenza dal fattore campo. Germania e Francia mostrano una differenza meno pronunciata, suggerendo una maggiore competitività anche fuori casa. Questa visualizzazione è particolarmente utile per valutare il vantaggio di giocare sul proprio terreno rispetto alle prestazioni esterne, fornendo un'intuizione su dove ciascuna squadra possa avere un vantaggio competitivo.

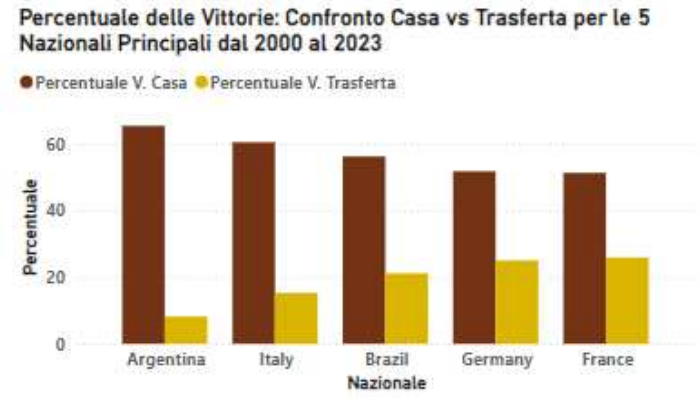


Figura 5.14: Grafico a colonne riguardante la percentuale delle vittorie in casa e in trasferta

5.3.3 Report sull'analisi dei goal

Nella Figura 5.15 abbiamo un grafico a ciambella che illustra i goal totali segnati dalle cinque squadre, con ogni segmento colorato per rappresentare una nazione. La dimensione

di ciascun segmento riflette la quantità di goal segnati, il che offre un colpo d'occhio sulla potenza offensiva di ciascuna squadra. L'inclusione delle percentuali aiuta a comprendere il contributo relativo di ciascuna squadra al conteggio complessivo dei goal. La distribuzione dei goal può riflettere le differenti strategie offensive e l'efficacia delle squadre durante il periodo analizzato.

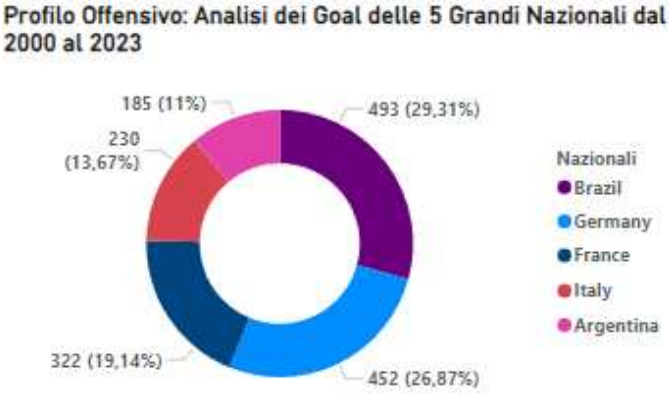


Figura 5.15: Grafico a ciambella riguardante i goal segnati dalle 5 nazionali principali

5.3.4 Report sull'analisi temporale dei goal

Nella Figura 5.16 esaminiamo un grafico a linee che mostra l'andamento temporale dei goal segnati dalle cinque nazionali principali dal 2000 al 2023. Le linee mostrano variazioni annuali significative nei goal segnati da ogni nazionale. Francia e Brasile presentano picchi evidenti in alcuni periodi, indicando anni particolarmente prolifici in termini di goal. Italia e Argentina mostrano un'andamento più stabile, con meno fluttuazioni estreme. La presenza di picchi e valli può essere correlata a competizioni specifiche, come i campionati mondiali o continentali, come pure a cambiamenti nelle formazioni e nelle strategie delle squadre. Le linee colorate permettono di tracciare le prestazioni storiche e possono indicare periodi di dominanza o transizioni nelle strategie di gioco. Fluttuazioni significative potrebbero essere collegate a eventi storici, a cambi di allenatori, o all'emergere di giocatori chiave.

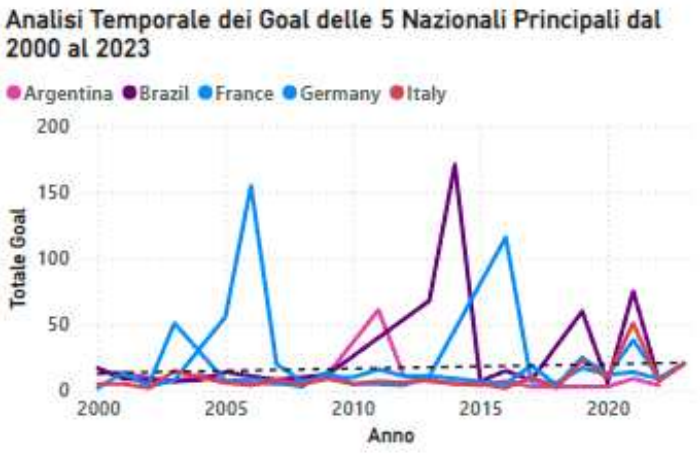


Figura 5.16: Grafico a linee riguardante l'andamento temporale dei goal segnati

Conclusioni e uno sguardo al futuro

Il percorso intrapreso in questa tesi, incentrato sulla realizzazione di una campagna di Data Analytics relativa ai risultati delle partite di calcio delle squadre nazionali, ha consentito di esporre e approfondire complesse questioni riguardanti la gestione e l'analisi dei Big Data, illustrando le potenzialità di tale approccio attraverso l'uso di Power BI.

Siamo partiti offrendo una dettagliata panoramica sulla Data Analytics, tracciando il quadro generale dei Big Data e passando dalla teoria delle 3V di Laney a quella delle 5V, descrivendo le categorie di Data Analytics e coprendo l'intero ciclo di vita della Big Data Analytics. Successivamente, abbiamo esaminato Power BI, mettendone in risalto le caratteristiche distintive e abbiamo approfondito il processo di ETL (Extract, Transform, Load), fondamentale nel contesto del data warehousing. Infine, ci siamo dedicati all'analisi generale e specifica dei dati, prendendo in considerazione le tabelle del dataset sui risultati delle partite delle squadre nazionali. Per queste analisi, grazie all'uso di misure e filtri, sono state realizzate quattro dashboard, una di carattere generale per prendere familiarità con il software, le altre tre più specifiche, riguardanti le tabelle del dataset.

Per quanto riguarda i possibili sviluppi futuri, questa tesi apre a diverse e interessanti prospettive di approfondimento e ampliamento. Nella gestione e nell'analisi dei Big Data, sarebbe utile esplorare ulteriormente le potenzialità offerte dall'integrazione di altre tecnologie avanzate, come il machine learning e l'Intelligenza Artificiale. L'implementazione di modelli predittivi potrebbe fornire ulteriori insight e favorire decisioni strategiche ancora più precise e informate. Un altro aspetto rilevante è l'espansione dell'analisi a dataset ancora più complessi e variegati, provenienti da fonti eterogenee. Questo permetterebbe di testare la scalabilità e la flessibilità delle metodologie e degli strumenti utilizzati, nonché di migliorare le capacità di integrazione e armonizzazione dei dati. Inoltre, un focus maggiore sulla visualizzazione avanzata dei dati potrebbe offrire ulteriori benefici, migliorando la comprensione e la comunicazione dei risultati analitici. Sperimentare con tecniche di visualizzazione interattiva e storytelling data-driven potrebbe rendere le dashboard ancora più efficaci e intuitive per gli utenti finali.

Infine, un'area di sviluppo importante riguarda l'ottimizzazione dei processi di ETL. Investire in soluzioni che automatizzino e migliorino l'efficienza di questi processi potrebbe ridurre significativamente i tempi di elaborazione e aumentare la qualità dei dati disponibili per l'analisi. Questi sviluppi futuri, combinati con una costante attenzione alle evoluzioni normative e alle tecniche del settore, contribuiranno a consolidare ulteriormente le pratiche di Data Analytics, rendendole sempre più indispensabili per il successo e la competitività delle aziende.

Bibliografia

CHEN, H., CHIANG, R. H. e STOREY, V. C. (2012), «Business Intelligence and Analytics: From Big Data to Big Impact», *MIS Quarterly*.

DEVLIN, B. (2013), *Business unIntelligence: Insight and Innovation beyond Analytics and Big Data*, Technics Publications, LLC, Basking Ridge, NJ.

DI NUZZO, M. (2021), *Data Science e Machine Learning: Dai dati alla conoscenza*, Michele di Nuzzo.

DRESNER, H. (2017), *2017 Big Data Analytics Market Study Report*, CreateSpace Independent Publishing Platform, Nashua, New Hampshire.

EMC EDUCATION SERVICES (2015), *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*, John Wiley & Sons, Hoboken, New Jersey.

FERRARI, A. e RUSSO, M. (2016), *Introducing Microsoft Power BI*, Microsoft Press, Redmond, Washington.

KNAFLIC, C. N. (2015), *Storytelling with Data: A Data Visualization Guide for Business Professionals*, John Wiley & Sons, Hoboken, New Jersey.

MACHIRAJU, S. e GAURAV, S. (2018), *Power BI Data Analysis and Visualization*, De Gruyter.

POWELL, B. (2018), *Mastering Microsoft Power BI: Expert techniques for effective data analytics and business intelligence*, Packt Publishing, Birmingham, UK.

PROVOST, F. e FAWCETT, T. (2013), *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*, O'Reilly Media, Sebastopol, CA.

RUSSO, M. e FERRARI, A. (2015), *The Definitive Guide to DAX: Business Intelligence with Microsoft Excel, SQL Server Analysis Services, and Power BI*, Microsoft Press, Redmond, Washington.

SALVAGGIO, A. (2023), *Business Intelligence con Microsoft Power BI: Guida completa per lanalisi e la visualizzazione dei dati*, Edizioni LSWR, Milano, Italia.

SHARDA, R., DELEN, D. e TURBAN, E. (2019), *Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support*, Pearson, Hoboken, New Jersey, 11th ed.

SPECTOR, A. Z., NORVIG, P., WIGGINS, C. e WING, J. M. (2023), *Data Science in Context: Foundations, Challenges, Opportunities*, Cambridge University Press, Cambridge, United Kingdom.

WEXLER, S., SHAFFER, J. e COTGREAVE, A. (2017), *The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios*, John Wiley & Sons, Hoboken, New Jersey.

Sitografia

- Kaggle – https://www.kaggle.com/
- Microsoft Power BI – https://learn.microsoft.com/en-us/power-bi/
- FIFA Official Website – https://www.fifa.com/it
- UEFA Official Website – https://www.uefa.com/
- Transfermarkt – https://www.transfermarkt.com/

Ringraziamenti

Desidero manifestare la mia profonda riconoscenza a tutti coloro che hanno contribuito a rendere il mio percorso universitario un'esperienza intensa e appagante.

Prima di tutto, vorrei ringraziare il mio relatore, il Professore Domenico Ursino per la sua instancabile dedizione, la sua pazienza e il suo sostegno costante durante questo percorso. Le sue importanti indicazioni e i suoi saggi consigli sono stati fondamentali per la realizzazione di questo lavoro.

Un particolare ringraziamento va anche al Dott. Luca Virgili, il mio correlatore, per le sue indicazioni e per il sostegno offerto durante la stesura di questa tesi.

Desidero manifestare la mia profonda gratitudine ai miei genitori e a mia sorella Federica, per il loro immenso affetto e per avermi costantemente sostenuto in tutte le mie scelte. Il vostro amore e supporto sono stati essenziali in questo lungo cammino.

Rivolgo un sentito ringraziamento ai miei zii, e in modo particolare alla cara zia Patrizia, per la sua costante presenza e per il suo incoraggiamento che mi ha sostenuto nel raggiungimento di questo importante obiettivo.

Un ringraziamento speciale va alla mia compagna di studi Isabella, con la quale ho condiviso tante giornate di intenso studio e anche momenti di spensieratezza, che mi hanno permesso di superare i vari ostacoli e raggiungere questo traguardo.

Vorrei ringraziare, inoltre, i miei amici, in particolare Francesco, Luigi, Roberto e Carlo. Il vostro sostegno, le vostre risate e i vostri consigli mi hanno permesso di affrontare e superare i momenti più difficili.

Grazie ai miei amici d'infanzia Vincenzo, Federico e Domenico che, nonostante la lontananza, mi hanno sempre supportato e regalato momenti di leggerezza.

Infine, desidero esprimere la mia immensa gratitudine ai miei nonni. Il vostro amore e il vostro supporto sono stati una guida lungo questo percorso carico di difficoltà.