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 science su dati riguardanti le malattie neurologiche

Design and implementation of a data science campaign on data concerning neurologic diseases

Relatore

Prof. Domenico Ursino

Correlatore

Dott. Michele Marchetti

Candidato

Samuel Ebibe Iwendi

ANNO ACCADEMICO 2025-2026

We are drowning in information
but starved for knowledge.

John Naisbitt

Sommario

Il presente lavoro di tesi si focalizza sull'applicazione di strumenti di Business Intelligence e Data Analytics in ambito neurofisiologico, con l'obiettivo di esplorare biomarcatori elettroencefalografici (EEG) su un campione clinico di 941 pazienti psichiatrici. Superando il tradizionale approccio basato sulla reportistica statica, è stato sviluppato un ecosistema analitico end-to-end all'interno della piattaforma Microsoft Power BI. Le attività principali hanno compreso un'estesa fase di Data Engineering (ETL) per la strutturazione dei dati spettrali multicanale e la progettazione di sei dashboard interattive. Attraverso il linguaggio DAX e l'integrazione di algoritmi di Machine Learning nativi (Clustering K-Means e Regressione Logistica), sono state indagate metriche avanzate – tra cui il Global Slowing Index, la Frontal Alpha Asymmetry e il Rapporto Theta/Beta – permettendo l'identificazione di pattern patologici e fenotipi trasversali rispetto alle diagnosi nosografiche tradizionali.

Keyword: Business Intelligence, Data Analytics, Elettroencefalogramma (EEG), Power BI, Explainable AI, Biomarcatori Neurofisiologici, Machine Learning, Psichiatria

Indice

Introduzione **1**

1 Introduzione alla Data Analytics **3**
1.1 Big Data . . . 3
1.1.1 Definizione e contesto storico . . . 3
1.1.2 Settori di applicazione . . . 4
1.1.3 Big Data nel settore sanitario . . . 4
1.2 Caratteristiche dei Big Data . . . 5
1.2.1 Volume . . . 6
1.2.2 Velocità . . . 6
1.2.3 Varietà . . . 6
1.2.4 Veracità . . . 7
1.2.5 Valore . . . 7
1.3 Data Analytics . . . 7
1.3.1 Definizione e inquadramento concettuale . . . 7
1.3.2 Il ruolo della Data Analytics nel supporto decisionale . . . 8
1.3.3 Data Analytics nel settore sanitario . . . 8
1.4 Categorie di Data Analytics . . . 8
1.4.1 Analisi Descrittiva . . . 9
1.4.2 Analisi Diagnostica . . . 10
1.4.3 Analisi Predittiva . . . 10
1.4.4 Analisi Prescrittiva . . . 11
1.5 Differenze tra Data Analytics e Data Analysis . . . 11
1.6 Ciclo di vita della Big Data Analytics . . . 12
1.6.1 Business Case Evaluation . . . 13
1.6.2 Data Identification . . . 14
1.6.3 Data Acquisition and Filtering . . . 14
1.6.4 Data Extraction . . . 14
1.6.5 Data Validation and Cleansing . . . 14
1.6.6 Data Aggregation and Representation . . . 15
1.6.7 Data Analysis . . . 15
1.6.8 Data Visualization . . . 17
1.6.9 Utilization of Analysis Results . . . 18

2 Introduzione a Microsoft Power BI **19**
2.1 Che cos'è Power BI . . . 19
2.1.1 Definizione e panoramica della piattaforma . . . 19
2.1.2 Posizionamento nel mercato della Business Intelligence . . . 20
2.1.3 Confronto con altre piattaforme . . . 20
2.2 Architettura e componenti principali . . . 21
2.2.1 Power BI Desktop . . . 21
2.2.2 Power BI Service . . . 22
2.2.3 Power BI Mobile . . . 22
2.3 Il modello dei dati in Power BI . . . 22
2.3.1 Il concetto di modello relazionale . . . 22
2.3.2 Lo schema a stella . . . 23
2.3.3 Gestione delle relazioni . . . 24
2.4 Le visualizzazioni in Power BI . . . 24
2.4.1 Tipologie di oggetti visivi . . . 24
2.4.2 Report e dashboard . . . 25
2.4.3 Filtri, slicer e interattività . . . 25
2.5 I linguaggi integrati: M e DAX . . . 26
2.5.1 Il linguaggio M e Power Query . . . 26
2.5.2 Il linguaggio DAX . . . 27
2.5.3 Il concetto di contesto in DAX . . . 27

3 Descrizione dei dati di partenza e attività di ETL **28**
3.1 L'elettroencefalogramma: cenni sui dati . . . 28
3.2 Il dataset clinico . . . 30
3.3 Il processo di ETL: inquadramento concettuale . . . 31
3.4 Fase di estrazione . . . 32
3.5 Fase di trasformazione e pulizia . . . 33
3.5.1 Trasformazioni in Power Query . . . 34
3.5.2 Feature Engineering in DAX: razionale tecnico . . . 34
3.5.3 Metriche di base (Misure) . . . 35
3.5.4 Potenza globale e normalizzazione (Colonne Calcolate) . . . 35
3.5.5 Biomarcatori sintetici e topologici . . . 36
3.6 Fase di caricamento e modellazione . . . 37

4 Progettazione e realizzazione delle dashboard di base **39**
4.1 Data Visualization e Business Intelligence nell'analisi clinica . . . 39
4.1.1 Principi cognitivi della visualizzazione dei dati . . . 39
4.1.2 Power BI come strumento di visualizzazione clinica . . . 40
4.2 Gli oggetti visivi scelti per l'analisi neurofisiologica . . . 40
4.3 Dashboard 1: Profilazione Clinica e Demografica . . . 41
4.3.1 Obiettivo e contesto analitico . . . 42
4.3.2 Descrizione degli oggetti visivi . . . 42
4.3.3 Risultati e interpretazione . . . 42
4.4 Dashboard 2: Firme Spettrali e Global EEG Slowing . . . 43
4.4.1 Fondamento teorico: il rallentamento globale dell'EEG . . . 43
4.4.2 Descrizione degli oggetti visivi . . . 44
4.4.3 Risultati e interpretazione . . . 44
4.5 Dashboard 3: Asse Cognitivo (Efficienza Neurale e Sincronizzazione Gamma) 45
4.5.1 Fondamento teorico: efficienza neurale e sincronizzazione Gamma . . . 45
4.5.2 Descrizione degli oggetti visivi . . . 46

4.5.3 Risultati e interpretazione 46
4.6 Dashboard 4: Asimmetria Alpha Frontale (FAA) 47
4.6.1 Fondamento teorico: la FAA e i modelli BIS/BAS 47
4.6.2 Descrizione degli oggetti visivi 48
4.6.3 Risultati e interpretazione 49

5 Progettazione e realizzazione delle dashboard avanzate **50**
5.1 Analisi avanzata in Power BI: strumenti e approcci 50
5.1.1 Dall'analisi diagnostica all'analisi predittiva 50
5.1.2 Gli strumenti nativi di Intelligenza Artificiale in Power BI 51
5.2 L'analisi predittiva nel contesto della diagnostica neurologica 51
5.3 Dashboard 5: Vulnerabilità allo stress e controllo esecutivo 52
5.3.1 Fondamento teorico: il Rapporto Theta/Beta (TBR) 53
5.3.2 Descrizione degli oggetti visivi 53
5.3.3 Risultati e interpretazione 54
5.4 Dashboard 6: Integrazione di modelli di Intelligenza Artificiale 55
5.4.1 Motore di selezione dinamica e architettura della dashboard 55
5.4.2 Apprendimento non supervisionato: Clustering K-Means 56
5.4.3 Apprendimento supervisionato: Explainable AI 56
5.4.4 Analisi della varianza: Decomposition Tree 57
5.4.5 Il loop analitico: integrazione e validazione incrociata 57

6 Discussione **59**
6.1 Discussione dei risultati 59
6.1.1 Il Global Slowing Index come discriminante patologico trasversale . . 59
6.1.2 Efficienza neurale e sincronizzazione Gamma: la conferma del gradiente 60
6.1.3 L'asimmetria Alpha Frontale e il dimorfismo sessuale nell'OCD 60
6.1.4 Il Theta/Beta Ratio e la "crisi di mezza età" corticale 60
6.1.5 Il Clustering K-Means e la fenotipizzazione trasversale 61
6.2 Punti di forza e limitazioni dello studio 61
6.2.1 Punti di forza 61
6.2.2 Limitazioni 61

Conclusioni **63**

Bibliografia **64**

Sitografia **67**

Ringraziamenti **68**

Elenco delle figure

1.1 Rappresentazione grafica del modello delle 5V dei *Big Data*: volume, velocità, varietà, veracità e valore 5
1.2 Le quattro categorie della *Data Analytics*: descrittiva, diagnostica, predittiva e prescrittiva, disposte lungo un asse di complessità crescente e valore informativo 9
1.3 Rappresentazione della relazione gerarchica tra *Data Analytics* e *Data Analysis*: la seconda costituisce un sottoinsieme della prima 12
1.4 Diagramma del ciclo di vita della *Big Data Analytics* secondo il *framework* di Erl, Khattak e Buhler; in esso sono evidenziate le nove fasi del processo, dalla valutazione del caso di studio all'utilizzazione dei risultati 13
1.5 Esempio schematico del processo di estrazione dei dati da fonti eterogenee . 15
1.6 Esempio di aggregazione di dati: due dataset vengono combinati utilizzando un campo chiave comune per creare una struttura dati unificata 16
1.7 Principali tipologie di visualizzazione dei dati classificate per tipo di informazione rappresentata: composizione, distribuzione, relazione e confronto . . . 17

2.1 Posizionamento delle principali piattaforme di *Business Intelligence* nel *Magic Quadrant* di Gartner (2026). Microsoft Power BI si colloca nel quadrante dei *Leader* 20
2.2 Architettura dell'ecosistema Power BI: i tre componenti principali – Desktop, Service e Mobile – e le loro interazioni 21
2.3 Rappresentazione schematica dello schema a stella: una tabella dei fatti centrale collegata a più tabelle dimensionali 23
2.4 Panoramica delle principali tipologie di oggetti visivi disponibili in Power BI, organizzate per categoria funzionale 25

3.1 Rappresentazione schematica del sistema internazionale 10-20 per il posizionamento degli elettrodi EEG. 29
3.2 Schermata iniziale di Power BI Desktop con le opzioni di importazione di dati. 32
3.3 Anteprima del dataset grezzo importato in Power BI, con le colonne anagrafiche, cliniche e le prime variabili spettrali della banda Delta. 33
3.4 Schermata di Power BI durante la creazione della colonna calcolata *Fascia di Istruzione*, con la formula DAX visibile nella barra delle formule e il pannello Dati sulla destra. 34

4.1 Panoramica di diversi tipi di Data Visualization. 40

4.2 Schermata della Dashboard 1 relativa alla profilazione clinica e demografica della coorte. 41
4.3 Schermata della Dashboard 2 relativa all'estrazione delle firme spettrali e al Global EEG Slowing. 43
4.4 Schermata della Dashboard 3 relativa all'Asse Cognitivo. 45
4.5 Schermata della Dashboard 4 relativa ai biomarcatori dell'umore e all'Asimmetria Alpha Frontale. 47

5.1 Schermata della Dashboard 5 relativa al Rapporto Theta/Beta e alla vulnerabilità allo stress esecutivo. 52
5.2 Schermata della Dashboard 6 relativa all'integrazione di modelli di Intelligenza Artificiale per la fenotipizzazione predittiva. 55

Introduzione

Le malattie neurologiche e psichiatriche rappresentano oggi una delle sfide sanitarie più urgenti a livello mondiale. Secondo l'Organizzazione Mondiale della Sanità, i disturbi mentali colpiscono quasi un miliardo di persone nel mondo e costituiscono una delle principali cause di disabilità e di perdita di anni di vita in buona salute. In questo scenario, la capacità di formulare diagnosi tempestive, accurate e fondate su evidenze oggettive diventa un imperativo clinico e scientifico.

L'elettroencefalografia (EEG), grazie alla sua elevata risoluzione temporale e alla sua natura non invasiva, si è affermata come uno degli strumenti più preziosi per l'esplorazione funzionale del cervello umano. L'analisi quantitativa del segnale EEG consente di estrarre biomarcatori spettrali — come le potenze delle bande Delta, Theta, Alpha, Beta e Gamma — che riflettono lo stato funzionale della corteccia cerebrale e che sono stati correlati, dalla letteratura scientifica, a specifiche condizioni patologiche. Tuttavia, la mole e la complessità di questi dati rendono la loro interpretazione un compito arduo: un singolo esame EEG a 19 elettrodi produce 95 valori di potenza spettrale, e un dataset clinico di centinaia di pazienti genera rapidamente matrici di dati la cui analisi manuale risulta impraticabile.

È in questo contesto che la *Data Analytics* e gli strumenti di *Business Intelligence* possono svolgere un ruolo decisivo. Tradizionalmente confinati al mondo aziendale per l'analisi di dati commerciali e finanziari, questi strumenti offrono capacità di trasformazione, modellazione e visualizzazione interattiva dei dati che si prestano naturalmente all'esplorazione di dataset biomedici complessi. La domanda di ricerca che anima il presente lavoro di tesi è dunque la seguente: *è possibile utilizzare una piattaforma di Business Intelligence come Microsoft Power BI non solo per descrivere, ma per scoprire pattern clinicamente rilevanti all'interno di dati elettroencefalografici, integrando tecniche di analisi descrittiva, diagnostica e predittiva in un unico ecosistema interattivo e trasparente?*

Per rispondere a questa domanda, il lavoro si è articolato lungo un percorso che attraversa l'intero ciclo di vita della *Big Data Analytics*: dalla definizione del problema clinico all'acquisizione e alla preparazione dei dati, dalla progettazione di biomarcatori derivati alla costruzione di dashboard analitiche progressivamente più complesse, fino all'integrazione di algoritmi di Intelligenza Artificiale nativi della piattaforma.

Il dataset alla base dello studio è costituito dalle registrazioni elettroencefalografiche di 941 pazienti affetti da un ampio spettro di disturbi psichiatrici — tra cui schizofrenia, disturbi dell'umore, disturbi d'ansia, disturbi ossessivo-compulsivi e dipendenze — oltre a un gruppo di controllo costituito da persone sane. Per ciascun paziente, il dataset riporta le potenze spettrali delle cinque bande di frequenza misurate su 19 elettrodi, unitamente a variabili demografiche e cliniche. A partire da questa matrice grezza, sono stati progettati e calcolati

biomarcatori neurofisiologici validati dalla letteratura internazionale, quali il *Global Slowing Index* (GSI), il *Frontal Alpha Asymmetry* (FAA) e il *Theta/Beta Ratio* (TBR), ciascuno dei quali cattura una dimensione specifica della disfunzione cerebrale.

Il cuore operativo del progetto è rappresentato da sei dashboard interattive, organizzate in due livelli di complessità crescente. Le prime quattro dashboard (analisi di base) coprono la profilazione demografica e clinica del campione, l'analisi del rallentamento corticale globale, lo studio dell'efficienza neurale e della sincronizzazione Gamma, e l'indagine sulle asimmetrie emotive frontali. Le ultime due dashboard (analisi avanzata) introducono il Rapporto Theta/Beta come indicatore di vulnerabilità allo stress esecutivo e, infine, integrano tre algoritmi di Intelligenza Artificiale — K-Means, regressione logistica e albero di scomposizione — in un unico pannello a filtraggio incrociato, capace di generare classificazioni predittive e di identificare fenotipi neuro-cognitivi latenti.

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

- Nel Capitolo 1 vengono introdotti i concetti fondamentali di *Big Data* e *Data Analytics*, con particolare attenzione alle applicazioni nel settore sanitario e neurologico. Si descrivono le cinque caratteristiche fondamentali dei *Big Data* (le cosiddette "5V"), le quattro categorie della *Data Analytics* e il ciclo di vita completo di un progetto di analisi dei dati.

- Nel Capitolo 2 viene presentato Microsoft Power BI, lo strumento di *Business Intelligence* scelto per la realizzazione del progetto. Si descrivono l'architettura della piattaforma, il modello di dati a stella, le principali tipologie di visualizzazione e i linguaggi integrati M e DAX.

- Nel Capitolo 3 si illustrano il dataset clinico di partenza e l'intero processo di ETL (*Extraction, Transformation, Loading*). Si forniscono le basi neurofisiologiche dell'elettroencefalografia, si descrive la struttura del dataset e si documentano rigorosamente tutte le trasformazioni effettuate in Power Query nonché le misure calcolate in DAX.

- Nel Capitolo 4 vengono progettate e analizzate le quattro dashboard di base, precedute da un inquadramento teorico sulla *Data Visualization* in ambito biomedico. Per ciascuna dashboard si descrivono il fondamento neurobiologico, l'architettura degli oggetti visivi e i risultati emersi dal campione clinico.

- Nel Capitolo 5 si presentano le due dashboard avanzate. Dopo un'introduzione teorica sull'analisi predittiva e sugli strumenti nativi di Intelligenza Artificiale in Power BI, si analizzano la dashboard dedicata al Rapporto Theta/Beta e le dashboard di integrazione dei modelli di *Machine Learning*.

- Nel Capitolo 6 vengono discussi criticamente i risultati alla luce della letteratura scientifica e si delineano i punti di forza e le limitazioni dello studio.

CAPITOLO 1

Introduzione alla Data Analytics

In questo capitolo viene introdotto il concetto di Data Analytics a partire dai Big Data, di cui si illustrano le principali caratteristiche, le diverse tipologie e le potenzialità applicative nei molteplici settori, con particolare riferimento all'ambito sanitario e alla ricerca sulle malattie neurologiche. Ampio spazio è dedicato alle cosiddette "5V" – volume, velocità, varietà, veracità e valore – che rappresentano le sfide e le opportunità nella gestione e nell'elaborazione di grandi moli di dati, quali quelli derivanti da registrazioni elettroencefalografiche e cartelle cliniche digitali. Successivamente, si approfondisce il ruolo strategico della Data Analytics come strumento essenziale per trasformare quantità ingenti di dati grezzi in conoscenza utile, a supporto dei processi decisionali e dell'innovazione in ambito clinico e scientifico. Il capitolo si conclude con una panoramica sulle quattro principali tipologie di Data Analytics, chiarendo la distinzione rispetto alla Data Analysis e presentando il ciclo di vita della Big Data Analytics.

1.1 Big Data

1.1.1 Definizione e contesto storico

Il concetto di *Big Data* ha assunto negli ultimi decenni una rilevanza crescente nel panorama scientifico, industriale e istituzionale, configurandosi come uno dei paradigmi più significativi dell'era digitale. Con il termine *Big Data* ci riferiamo a insiemi di dati la cui dimensione, complessità e velocità di generazione superano le capacità dei tradizionali strumenti di gestione e analisi dei dati. Secondo la definizione proposta da Gartner, i *Big Data* rappresentano «risorse informative ad alto volume, alta velocità e/o alta varietà che richiedono forme innovative ed economicamente sostenibili di elaborazione delle informazioni, al fine di consentire una comprensione approfondita, il supporto alle decisioni e l'automazione dei processi» [Gartner, 2012].

Per comprendere appieno la portata di tale fenomeno, è opportuno ripercorrere brevemente l'evoluzione storica della gestione dei dati. Fino alla fine del XX secolo, la raccolta, l'archiviazione e l'elaborazione delle informazioni avvenivano prevalentemente attraverso *database* relazionali e sistemi strutturati, progettati per gestire volumi di dati relativamente contenuti e caratterizzati da schemi rigidi e predefiniti. I sistemi di gestione di basi di dati relazionali, introdotti a partire dagli anni Settanta, hanno rappresentato per diversi decenni lo standard di riferimento per l'organizzazione e l'interrogazione dei dati [Codd, 1970]. Tali sistemi si dimostravano adeguati in contesti in cui i dati erano prevalentemente strutturati,

generati a velocità moderate e conservati in quantità gestibili con le risorse computazionali disponibili.

La situazione è mutata radicalmente con l'avvento della rivoluzione digitale. A partire dalla fine degli anni Novanta e, in modo ancora più marcato, nel corso del primo decennio del XXI secolo, una serie di fenomeni tecnologici e sociali ha determinato una crescita esponenziale nella produzione di dati. La diffusione capillare di Internet, l'espansione dei *social media*, la proliferazione dei dispositivi mobili e, più recentemente, lo sviluppo dell'*Internet of Things* (IoT) hanno generato flussi di dati senza precedenti nella storia dell'umanità. Secondo alcune stime, nel 2020 la quantità di dati generati quotidianamente a livello globale ha superato i 2,5 quintilioni di *byte*, una cifra destinata a crescere ulteriormente nei prossimi anni [Domo, 2020].

Questa esplosione nella produzione dei dati non riguarda soltanto l'incremento quantitativo, ma coinvolge anche la natura stessa delle informazioni generate. Accanto ai dati strutturati tradizionali – tabelle, record numerici, transazioni finanziarie – si è assistito alla comparsa massiccia di dati semi-strutturati e non strutturati: testi in linguaggio naturale, immagini, contenuti audio e video, segnali provenienti da sensori, tracce di navigazione e interazioni digitali di varia natura. Tale eterogeneità ha reso inadeguati gli strumenti tradizionali di gestione dei dati, stimolando la ricerca e lo sviluppo di nuove architetture e metodologie capaci di affrontare le sfide poste da questa nuova realtà informativa.

1.1.2 Settori di applicazione

Le potenzialità applicative dei *Big Data* si estendono a una molteplicità di settori. In ambito industriale, l'analisi di grandi volumi di dati consente di ottimizzare i processi produttivi, prevedere guasti nelle catene di produzione e personalizzare l'offerta commerciale in funzione delle preferenze dei consumatori [Manyika *et al.*, 2011]. Nel settore finanziario, i *Big Data* trovano applicazione nella valutazione del rischio creditizio, nella prevenzione delle frodi e nell'analisi in tempo reale dei mercati. Nell'ambito della Pubblica Amministrazione, l'impiego dei dati su larga scala permette di migliorare la pianificazione urbana, di monitorare i flussi migratori e di supportare le politiche sociali con evidenze empiriche. In campo scientifico, la disponibilità di enormi quantità di dati ha rivoluzionato discipline quali la genomica, l'astronomia e la fisica delle particelle, accelerando il ritmo delle scoperte e aprendo nuove frontiere di ricerca [Hey *et al.*, 2009].

1.1.3 Big Data nel settore sanitario

Tra i molteplici ambiti in cui i *Big Data* esercitano un impatto trasformativo, il settore sanitario riveste un'importanza del tutto particolare. La sanità moderna genera quantità enormi di dati provenienti da fonti eterogenee: cartelle cliniche elettroniche, referti di laboratorio, immagini diagnostiche, dati genomici, segnali biomedici e, in misura crescente, informazioni raccolte da dispositivi indossabili e applicazioni di telemedicina [Raghupathi e Raghupathi, 2014]. Si stima che il volume complessivo dei dati sanitari raddoppi ogni due o tre anni, una crescita che pone sfide considerevoli ma che offre, al contempo, opportunità straordinarie per il miglioramento della qualità delle cure e per l'avanzamento della conoscenza medica.

Nel contesto specifico della neurologia e della psichiatria, l'applicazione dei *Big Data* assume connotazioni particolarmente significative. Le malattie neurologiche e psichiatriche – tra cui l'epilessia, il morbo di Alzheimer, il disturbo depressivo maggiore, i disturbi d'ansia e le patologie cerebrovascolari – costituiscono una delle principali cause di disabilità a livello mondiale e rappresentano un onere crescente per i sistemi sanitari [World Health Organization, 2019]. La comprensione dei meccanismi fisiopatologici alla base di tali condizioni richiede l'analisi di dati complessi e multidimensionali, provenienti da esami strumentali

quali l'elettroencefalogramma, la risonanza magnetica funzionale e la tomografia a emissione di positroni [Gandomi e Haider, 2015].

La convergenza tra *Big Data* e neuroscienze apre prospettive di grande interesse: dall'identificazione di *biomarker* elettrofisiologici per la diagnosi precoce delle patologie neurodegenerative, alla stratificazione dei pazienti in sottogruppi clinicamente omogenei, fino alla personalizzazione dei percorsi terapeutici sulla base di profili individuali ricavati dall'analisi integrata di dati eterogenei. Affinché tali potenzialità possano tradursi in risultati concreti è necessario disporre di strumenti analitici adeguati e di metodologie rigorose per la gestione, la pulizia e l'interpretazione dei dati.

1.2 Caratteristiche dei Big Data

Per comprendere la natura e le implicazioni dei *Big Data*, la letteratura scientifica ha progressivamente elaborato un modello descrittivo fondato su un insieme di proprietà fondamentali, comunemente note come le "5V". Tale modello, inizialmente proposto da Doug Laney nel 2001 con riferimento a tre dimensioni – volume, velocità e varietà – [Laney, 2001], è stato successivamente ampliato con l'aggiunta di veracità e valore, al fine di catturare in modo più completo la complessità del fenomeno [Gandomi e Haider, 2015]. Nella Figura 1.1 riportiamo una rappresentazione grafica del modello delle 5V, che costituisce il quadro concettuale di riferimento per la trattazione che segue.

Figura 1.1: Rappresentazione grafica del modello delle 5V dei *Big Data*: volume, velocità, varietà, veracità e valore

1.2.1 Volume

La prima, e forse più intuitiva, caratteristica dei *Big Data* è il volume, ovvero la dimensione estremamente elevata degli insiemi di dati coinvolti. Il concetto di "grande volume" è intrinsecamente relativo e dipende dal contesto storico e tecnologico: ciò che in un dato momento viene considerato un volume di dati eccezionale può diventare ordinario nell'arco di pochi anni, in virtù dell'evoluzione delle capacità di calcolo e di archiviazione. Ciononostante, l'ordine di grandezza dei dati con cui ci confrontiamo oggi supera di gran lunga quello di qualsiasi epoca precedente. Si parla correntemente di *petabyte* (10^{15} *byte*), *exabyte* (10^{18} *byte*) e persino *zettabyte* (10^{21} *byte*) per descrivere i volumi complessivi di dati generati a livello globale [Hilbert e López, 2011].

Nel settore sanitario, il volume dei dati è alimentato da molteplici fonti. I sistemi informativi ospedalieri producono quotidianamente migliaia di record relativi a ricoveri, dimissioni, prescrizioni farmacologiche e risultati di esami di laboratorio. I sistemi di diagnostica per immagini generano file di dimensioni considerevoli: una singola risonanza magnetica cerebrale può occupare diverse centinaia di megabyte, mentre le immagini ad alta risoluzione prodotte dalla tomografia computerizzata raggiungono facilmente il gigabyte [Raghupathi e Raghupathi, 2014]. Anche i segnali biomedici, come le registrazioni elettroencefalografiche, producono serie temporali multicanale la cui dimensione complessiva, moltiplicata per ampie coorti di pazienti, raggiunge facilmente l'ordine dei gigabyte.

1.2.2 Velocità

La seconda caratteristica fondamentale dei *Big Data* è la velocità, intesa sia come rapidità con cui i dati vengono generati, sia come necessità di elaborarli in tempi compatibili con le esigenze applicative. In numerosi contesti, il valore dell'informazione è strettamente legato alla tempestività con cui essa viene resa disponibile: un dato analizzato con ritardo può perdere gran parte della propria utilità decisionale.

Nel settore sanitario, la dimensione della velocità assume un rilievo particolare. I sistemi di monitoraggio in tempo reale, quali quelli impiegati nelle unità di terapia intensiva, generano flussi continui di dati relativi ai parametri vitali dei pazienti – frequenza cardiaca, pressione arteriosa, saturazione dell'ossigeno – che devono essere analizzati istantaneamente per consentire interventi tempestivi in caso di anomalie. Analogamente, il monitoraggio elettroencefalografico continuo, utilizzato ad esempio nella diagnosi e nella gestione dell'epilessia, produce un flusso ininterrotto di dati che richiede l'applicazione di algoritmi capaci di operare in *near real-time*, al fine di identificare prontamente eventi critici, quali le crisi epilettiche [Abou Jaoude *et al.*, 2020]. Anche quando l'analisi avviene in modalità *batch* – ossia su dati già acquisiti e archiviati – la velocità con cui i risultati vengono prodotti e resi accessibili influenza direttamente l'utilità clinica e decisionale dell'analisi.

1.2.3 Varietà

La terza caratteristica dei *Big Data* è la varietà, che si riferisce alla molteplicità delle tipologie e dei formati con cui i dati si presentano. A differenza dei tradizionali *database* relazionali, che operano su dati strutturati e organizzati secondo schemi predefiniti, i *Big Data* comprendono una gamma estremamente eterogenea di informazioni.

I dati strutturati sono quelli organizzati in tabelle con righe e colonne, conformi a uno schema fisso; esempi tipici nel contesto sanitario sono le tabelle anagrafiche dei pazienti, i risultati numerici degli esami di laboratorio e i codici diagnostici secondo la classificazione internazionale delle malattie. I dati semi-strutturati possiedono una struttura parziale, ma non rigidamente definita, come nel caso dei referti clinici redatti in formato testuale ma

accompagnati da metadati strutturati. I dati non strutturati, infine, sono privi di un'organizzazione predefinita: immagini diagnostiche, segnali EEG grezzi, registrazioni audio di colloqui clinici e annotazioni testuali libere rientrano in questa categoria [Gandomi e Haider, 2015].

1.2.4 Veracità

La quarta dimensione del modello è la veracità, che riguarda la qualità, l'affidabilità e l'accuratezza dei dati. In un contesto caratterizzato da volumi enormi e fonti eterogenee, il rischio di incorrere in dati incompleti, incoerenti, erronei o fuorvianti è particolarmente elevato. La veracità rappresenta pertanto una sfida critica, poiché la qualità delle analisi e delle decisioni basate sui dati dipende in modo diretto dalla qualità dei dati stessi.

Nel settore sanitario, la questione della veracità assume una rilevanza ancora maggiore, in ragione delle potenziali conseguenze cliniche derivanti da analisi condotte su dati inaffidabili. Le cartelle cliniche elettroniche possono contenere errori di trascrizione, codifiche diagnostiche imprecise o informazioni incomplete. I segnali biomedici, quali le registrazioni elettroencefalografiche, sono soggetti ad artefatti di varia natura – movimenti oculari, attività muscolare, interferenze elettriche ambientali – che possono compromettere la qualità del dato se non adeguatamente identificati e trattati [Islam *et al.*, 2016]. Garantire la veracità dei dati sanitari è, dunque, una condizione necessaria, prima ancora che sufficiente, per produrre analisi clinicamente rilevanti e scientificamente valide.

1.2.5 Valore

La quinta e ultima dimensione del modello è il valore, che rappresenta, in ultima analisi, la ragione stessa per cui i *Big Data* vengono raccolti e analizzati. Il valore si riferisce alla capacità di estrarre informazioni utili, conoscenza azionabile e *insight* significativi dalla massa di dati grezzi disponibili. In altri termini, il valore dei dati non risiede nella loro mera esistenza, bensì nella possibilità di trasformarli in supporto concreto per i processi decisionali, per l'innovazione e per il miglioramento dei risultati in un dato contesto applicativo.

Nel settore sanitario, il valore dei *Big Data* si esprime nella possibilità di migliorare la qualità delle diagnosi, di prevedere l'evoluzione delle patologie, di personalizzare i trattamenti terapeutici e di ottimizzare l'allocazione delle risorse. Nel contesto delle neuroscienze cliniche, il valore dei dati biomedici risiede nella capacità di identificare *pattern* elettrofisiologici e clinici associati a specifiche condizioni patologiche, di stratificare i pazienti in sottogruppi omogenei e di fornire ai clinici strumenti di supporto decisionale basati su evidenze quantitative [Bzdok e Meyer-Lindenberg, 2018].

1.3 Data Analytics

1.3.1 Definizione e inquadramento concettuale

Con il termine *Data Analytics* intendiamo l'insieme dei processi, delle tecniche e degli strumenti volti all'ispezione, alla pulizia, alla trasformazione e alla modellazione dei dati, con l'obiettivo di scoprire informazioni utili, suggerire conclusioni e supportare i processi decisionali [Davenport e Harris, 2007]. La *Data Analytics* non si esaurisce nella semplice raccolta o archiviazione dei dati, ma comprende l'intero ciclo che va dall'acquisizione del dato grezzo alla generazione di conoscenza azionabile.

È opportuno distinguere la *Data Analytics* dalla mera raccolta di dati e dalla loro presentazione descrittiva. Mentre la raccolta dei dati rappresenta una fase necessaria ma non sufficiente, e la presentazione descrittiva si limita a fornire una fotografia dello stato corrente

delle informazioni disponibili, la *Data Analytics* si caratterizza per un approccio sistematico e metodologicamente rigoroso, finalizzato all'estrazione di *pattern*, alla formulazione di ipotesi e alla verifica di relazioni significative all'interno dei dati [Provost e Fawcett, 2013].

Il concetto di *Data Analytics* si colloca al crocevia di diverse discipline: la statistica, l'informatica, la matematica applicata e le scienze del dominio specifico in cui l'analisi viene condotta. Tale natura interdisciplinare rende la *Data Analytics* uno strumento estremamente versatile, applicabile a contesti tra loro molto diversi, dall'analisi dei mercati finanziari alla ricerca biomedica, dalla gestione delle catene di approvvigionamento alla comprensione dei fenomeni sociali.

1.3.2 Il ruolo della Data Analytics nel supporto decisionale

Uno degli aspetti più rilevanti della *Data Analytics* risiede nella sua capacità di supportare i processi decisionali in contesti caratterizzati da incertezza e complessità. In ambito organizzativo, i sistemi di supporto alle decisioni si avvalgono delle tecniche di *Data Analytics* per trasformare i dati disponibili in informazioni strutturate, sulla base delle quali i responsabili possono compiere scelte informate e tempestive [Power, 2002].

La transizione da un approccio decisionale basato sull'intuizione e sull'esperienza individuale a un approccio fondato sull'evidenza empirica rappresenta uno dei cambiamenti più significativi nella cultura organizzativa contemporanea. In numerosi settori, la capacità di analizzare i dati in modo efficace si è trasformata in un vantaggio competitivo determinante. Le organizzazioni in grado di sfruttare appieno il potenziale della *Data Analytics* possono anticipare le tendenze di mercato, ottimizzare i processi interni e rispondere con maggiore rapidità ai cambiamenti del contesto operativo [Davenport e Harris, 2007].

1.3.3 Data Analytics nel settore sanitario

L'applicazione della *Data Analytics* nel settore sanitario ha conosciuto una crescita significativa nel corso degli ultimi anni, in parallelo con la digitalizzazione progressiva dei sistemi informativi ospedalieri e con l'aumento esponenziale dei dati clinici disponibili. Le potenzialità della *Data Analytics* in ambito sanitario sono molteplici e comprendono, tra le altre, la predizione dell'insorgenza di patologie, la personalizzazione dei percorsi terapeutici, l'ottimizzazione della gestione delle risorse sanitarie e il miglioramento della qualità complessiva delle cure [Raghupathi e Raghupathi, 2014].

Nel campo della neurologia e delle neuroscienze cliniche, la *Data Analytics* offre strumenti particolarmente preziosi. I segnali biomedici cerebrali producono grandi volumi di dati complessi e multidimensionali, la cui interpretazione manuale da parte del clinico risulta dispendiosa in termini di tempo e soggetta a variabilità inter-osservatore. L'impiego di tecniche analitiche avanzate consente di automatizzare parzialmente il processo di interpretazione, di identificare *pattern* ricorrenti e di mettere in luce relazioni tra variabili cliniche che potrebbero sfuggire all'analisi tradizionale [Bzdok e Meyer-Lindenberg, 2018].

1.4 Categorie di Data Analytics

Dopo aver introdotto il concetto generale di *Data Analytics* e il suo ruolo nel processo decisionale, è opportuno approfondire le diverse tipologie di analisi che compongono questo ecosistema. La letteratura scientifica individua comunemente quattro categorie fondamentali, ciascuna caratterizzata da un livello crescente di complessità, di valore informativo e di difficoltà implementativa [Banerjee *et al.*, 2013]. Tali categorie rispondono a domande progressivamente più sofisticate: dall'esplorazione di ciò che è accaduto, alla comprensione

delle cause, fino alla previsione degli eventi futuri e alla formulazione di raccomandazioni operative.

Nella Figura 1.2 presentiamo uno schema riassuntivo delle quattro categorie, disposte lungo un asse che evidenzia la relazione tra complessità analitica e valore generato per l'organizzazione. Come possiamo osservare, il passaggio dall'analisi descrittiva a quella prescrittiva comporta un incremento significativo sia delle competenze tecniche richieste sia del potenziale impatto sul processo decisionale [Delen e Demirkan, 2013].

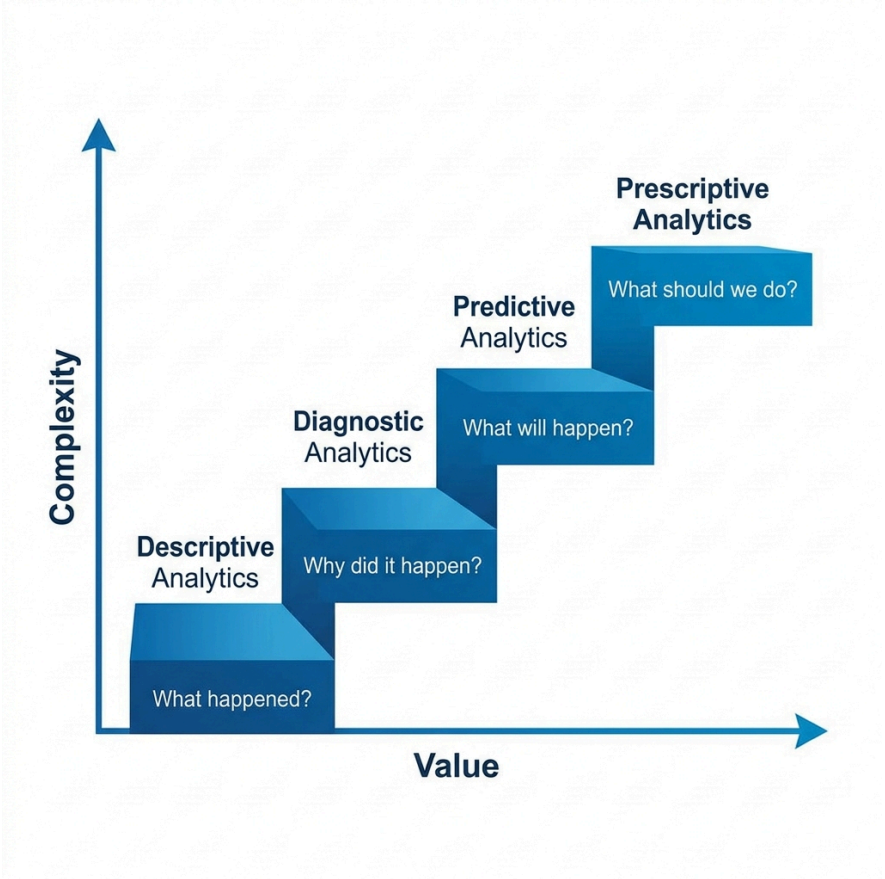


Figura 1.2: Le quattro categorie della *Data Analytics*: descrittiva, diagnostica, predittiva e prescrittiva, disposte lungo un asse di complessità crescente e valore informativo

Le quattro categorie non sono mutuamente esclusive, ma rappresentano stadi complementari di un percorso analitico integrato. Un progetto di *Data Analytics* maturo attraversa tipicamente più livelli: partendo dalla semplice descrizione dei dati storici, può spingersi fino alla previsione di eventi futuri e alla formulazione di raccomandazioni operative.

1.4.1 Analisi Descrittiva

L'analisi descrittiva rappresenta il primo e più diffuso livello di *Data Analytics*. Il suo obiettivo consiste nel rispondere alla domanda fondamentale: *cosa è successo?* Attraverso tecniche di sintesi statistica, aggregazione e visualizzazione, questa tipologia di analisi trasforma i dati grezzi in informazioni comprensibili, offrendo una fotografia dello stato attuale o storico di un fenomeno [Banerjee *et al.*, 2013].

Gli strumenti principali dell'analisi descrittiva comprendono:

- gli indicatori chiave di prestazione (KPI), che condensano grandezze complesse in metriche sintetiche e immediatamente interpretabili;

- le *dashboard* interattive, che aggregano molteplici visualizzazioni in un unico ambiente di consultazione;
- i *report* periodici, che documentano l'andamento di variabili rilevanti nel tempo;
- le statistiche riassuntive, quali medie, mediane, distribuzioni di frequenza e percentili.

L'analisi descrittiva non si propone di individuare relazioni causali né di formulare previsioni: il suo scopo è rendere accessibile e leggibile la realtà codificata nei dati. Nonostante questa apparente semplicità, essa costituisce il fondamento indispensabile su cui poggiano le forme di analisi più avanzate. Senza una comprensione solida di ciò che i dati raccontano, qualsiasi tentativo di diagnosi o previsione risulterebbe privo di basi.

1.4.2 Analisi Diagnostica

L'analisi diagnostica rappresenta il secondo livello e risponde alla domanda: *perché è successo?* Se l'analisi descrittiva ci informa su ciò che i dati registrano, quella diagnostica si spinge oltre, cercando di identificare le cause e i fattori che spiegano i fenomeni osservati [Delen e Demirkan, 2013].

Le tecniche caratteristiche di questa categoria includono:

- l'analisi di correlazione, che misura il grado di associazione tra variabili;
- il *drill-down*, ossia l'esplorazione progressiva dei dati a livelli di dettaglio crescente;
- la *root cause analysis*, un approccio sistematico per risalire alle cause originarie di un evento o di un'anomalia;
- la segmentazione, che suddivide la popolazione in sottogruppi omogenei per evidenziare differenze significative.

L'analisi diagnostica richiede generalmente un intervento più attivo dell'analista, il quale deve formulare ipotesi, selezionare le variabili rilevanti e interpretare i risultati alla luce del dominio applicativo. A differenza dell'analisi descrittiva, che può essere largamente automatizzata, quella diagnostica presuppone una conoscenza approfondita del contesto e la capacità di porre le domande giuste ai dati.

1.4.3 Analisi Predittiva

L'analisi predittiva costituisce il terzo livello e affronta la domanda: *cosa succederà?* Questa tipologia si distingue dalle precedenti perché non si limita a osservare e spiegare il passato, ma utilizza i dati storici per costruire modelli in grado di stimare eventi, tendenze o comportamenti futuri [Banerjee *et al.*, 2013].

Le tecniche e gli strumenti propri dell'analisi predittiva comprendono:

- i modelli di regressione statistica, che quantificano la relazione tra variabili predittive e una variabile di risposta;
- gli algoritmi di *machine learning*, quali alberi decisionali, reti neurali, macchine a vettori di supporto e metodi di insieme;
- le tecniche di *clustering*, come il *K-Means*, che raggruppano le osservazioni in insiemi omogenei sulla base di similarità intrinseche;
- i metodi di *forecasting*, che proiettano nel tempo l'andamento di serie temporali.

L'analisi predittiva non garantisce certezze, ma fornisce stime probabilistiche che, se costruite su dati di qualità e modelli adeguati, possono orientare efficacemente il processo decisionale. Il suo valore risiede nella capacità di anticipare scenari, consentendo alle organizzazioni di prepararsi e di allocare le risorse in modo proattivo anziché reattivo.

1.4.4 Analisi Prescrittiva

L'analisi prescrittiva rappresenta il livello più avanzato e complesso della *Data Analytics*. Essa risponde alla domanda: *cosa dovremmo fare?* Se l'analisi predittiva fornisce stime su ciò che potrebbe accadere, quella prescrittiva si spinge fino a suggerire le azioni ottimali da intraprendere per raggiungere un obiettivo desiderato o per mitigare un rischio identificato [Delen e Demirkan, 2013].

Le tecniche proprie di questa categoria includono:

- i modelli di ottimizzazione, che identificano la soluzione migliore in uno spazio di alternative vincolate;
- i sistemi di raccomandazione, che suggeriscono azioni personalizzate sulla base del profilo dell'utente o del paziente;
- la simulazione, che valuta gli effetti di scenari alternativi prima della loro implementazione;
- i sistemi di supporto alle decisioni, che integrano modelli analitici e conoscenza del dominio per assistere il decisore.

L'analisi prescrittiva richiede non soltanto una solida infrastruttura analitica, ma anche la capacità di tradurre i risultati dei modelli in raccomandazioni operative concrete. In ambito sanitario, ad esempio, un sistema prescrittivo potrebbe suggerire il protocollo terapeutico più appropriato per un paziente sulla base del suo profilo clinico, oppure potrebbe indicare la priorità con cui sottoporre i pazienti a ulteriori accertamenti diagnostici.

1.5 Differenze tra Data Analytics e Data Analysis

Nel linguaggio comune, e talvolta anche nella letteratura non specialistica, i termini *Data Analytics* e *Data Analysis* vengono utilizzati in modo intercambiabile, come se indicassero il medesimo concetto. In realtà, tra i due esiste una differenza sostanziale che è opportuno chiarire per evitare ambiguità terminologiche e per inquadrare correttamente il lavoro svolto in questa tesi.

Con l'espressione *Data Analysis* ci riferiamo a una fase specifica del processo analitico, quella in cui i dati vengono esaminati, trasformati e interpretati con l'obiettivo di estrarre informazioni utili. La *Data Analysis* è, dunque, un'attività circoscritta, che si concentra sull'applicazione di tecniche statistiche, algoritmi e metodi di esplorazione a un *dataset* definito. Essa include operazioni quali il calcolo di statistiche descrittive, la verifica di ipotesi, l'identificazione di *pattern* e la costruzione di modelli. Il suo perimetro è delimitato nel tempo e nello scopo: analizzare un insieme di dati per rispondere a una domanda specifica.

La *Data Analytics*, al contrario, designa un ecosistema molto più ampio e articolato. Essa comprende l'intero insieme di strumenti, metodologie, processi e tecnologie che un'organizzazione utilizza per trasformare i dati grezzi in conoscenza operativa. La *Data Analytics* abbraccia il ciclo di vita completo del dato: dalla sua identificazione e acquisizione, passando per la pulizia, l'integrazione e l'analisi vera e propria, fino alla visualizzazione dei risultati e al loro utilizzo per il processo decisionale [Erl *et al.*, 2016].

Possiamo dunque affermare che la *Data Analysis* è una componente, un sottoinsieme, della *Data Analytics*. Nella Figura 1.3 illustriamo questa relazione gerarchica attraverso un diagramma concettuale che evidenzia come la *Data Analysis* si collochi all'interno del più ampio ecosistema della *Data Analytics*, il quale include anche le fasi di acquisizione dei dati, di gestione dell'infrastruttura tecnologica, di visualizzazione e di comunicazione dei risultati.

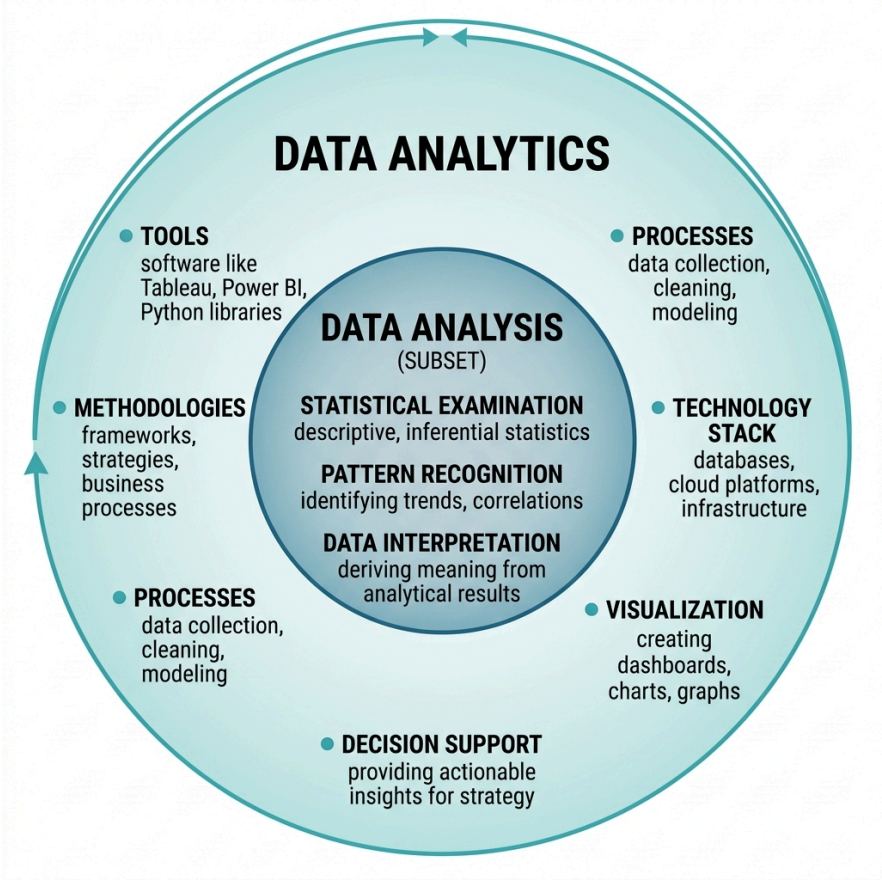


Figura 1.3: Rappresentazione della relazione gerarchica tra *Data Analytics* e *Data Analysis*: la seconda costituisce un sottoinsieme della prima

1.6 Ciclo di vita della Big Data Analytics

Per comprendere come un progetto di *Data Analytics* si articoli nella pratica, è utile fare riferimento a un modello di ciclo di vita che ne descriva le fasi in modo sistematico. In questa sezione adottiamo il *framework* proposto da Erl, Khattak e Buhler [Erl *et al.*, 2016], il quale suddivide il processo di *Big Data Analytics* in nove fasi sequenziali, ciascuna con obiettivi, attività e risultati ben definiti.

Nella Figura 1.4 presentiamo il diagramma del ciclo di vita, che illustra la successione delle nove fasi e le relazioni di dipendenza tra di esse. Come possiamo osservare, il processo non è strettamente lineare: alcune fasi possono richiedere iterazioni e ritorni a stadi precedenti, in particolare quando l'analisi dei dati rivela la necessità di acquisire nuove fonti o di raffinare le operazioni di pulizia.

Nelle sottosezioni che seguono, descriviamo ciascuna fase del ciclo di vita, illustrandone le caratteristiche principali e le problematiche tipiche che un analista si trova ad affrontare nella pratica.

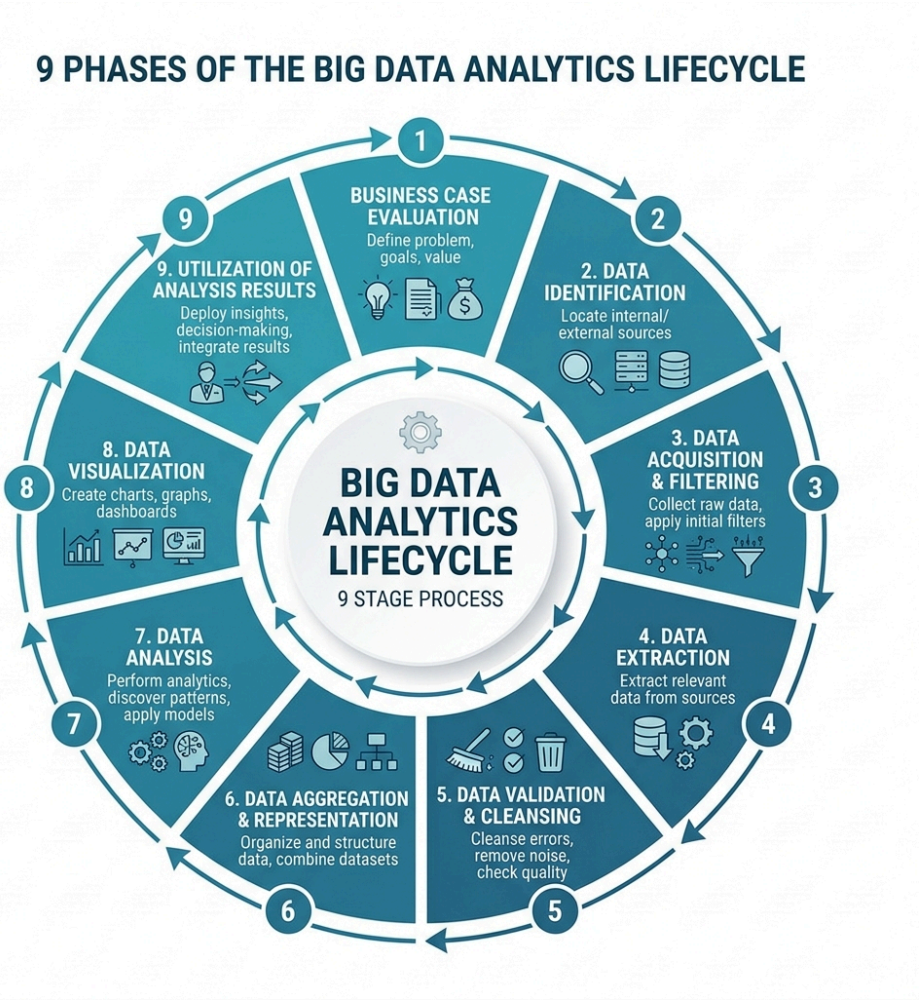


Figura 1.4: Diagramma del ciclo di vita della *Big Data Analytics* secondo il *framework* di Erl, Khattak e Buhler; in esso sono evidenziate le nove fasi del processo, dalla valutazione del caso di studio all'utilizzazione dei risultati

1.6.1 Business Case Evaluation

La prima fase del ciclo di vita consiste nella valutazione del caso di studio. Il suo scopo è definire con chiarezza il problema che si intende risolvere, identificare gli obiettivi dell'analisi e valutare la fattibilità del progetto in termini di risorse, competenze e dati disponibili [Erl *et al.*, 2016].

In questa fase, l'organizzazione o il gruppo di ricerca deve rispondere ad alcune domande fondamentali:

- Quale problema si intende affrontare e quale valore aggiunto può apportare l'analisi dei dati?
- Quali sono le domande specifiche a cui si desidera rispondere?
- I dati necessari sono disponibili oppure devono essere acquisiti?
- Le competenze tecniche e le risorse infrastrutturali sono adeguate?

La valutazione del caso di studio non è un esercizio formale, ma una fase critica che condiziona l'intero progetto. Un problema mal definito o obiettivi ambigui possono compromettere tutte le fasi successive, indipendentemente dalla qualità dei dati e degli strumenti utilizzati.

1.6.2 Data Identification

La seconda fase riguarda l'identificazione dei dati. Una volta definito il problema, è necessario individuare le fonti di dati che contengono le informazioni pertinenti alla risoluzione del caso di studio. Questa fase richiede una mappatura sistematica delle sorgenti disponibili, che possono includere basi di dati interne, archivi pubblici, flussi di dati in tempo reale o *dataset* prodotti da strumentazione scientifica [Erl *et al.*, 2016].

L'identificazione dei dati non si limita alla localizzazione delle fonti, ma comprende anche una valutazione preliminare della loro rilevanza, qualità e accessibilità. È opportuno verificare, in questa fase, se i dati sono disponibili nel formato desiderato, se coprono l'arco temporale di interesse e se la loro granularità è sufficiente per le analisi previste.

1.6.3 Data Acquisition and Filtering

La terza fase del ciclo di vita riguarda l'acquisizione e il filtraggio dei dati. Una volta identificate le fonti, i dati devono essere effettivamente ottenuti e sottoposti a un primo processo di selezione, che consente di eliminare le informazioni irrilevanti o ridondanti prima delle fasi di lavorazione successive [Erl *et al.*, 2016].

L'acquisizione può avvenire attraverso modalità diverse a seconda della natura della fonte: importazione di file, interrogazione di basi di dati, connessione a interfacce di programmazione (API) o raccolta di flussi di dati in tempo reale. Il filtraggio iniziale, dal canto suo, ha lo scopo di ridurre il volume dei dati a un sottoinsieme gestibile e pertinente, eliminando record duplicati, campi non necessari o osservazioni manifestamente errate.

1.6.4 Data Extraction

La quarta fase riguarda l'estrazione dei dati, ossia la trasformazione dei dati acquisiti in un formato compatibile con gli strumenti di analisi e visualizzazione che si intendono utilizzare. L'estrazione può comportare la conversione tra formati differenti, la ristrutturazione delle tabelle, la rinominazione dei campi o l'adattamento delle codifiche [Erl *et al.*, 2016].

In molti progetti di *Data Analytics*, questa fase rappresenta un passaggio critico, poiché i dati provenienti da fonti eterogenee raramente si presentano in una forma direttamente utilizzabile (come schematizzato nella Figura 1.5).

1.6.5 Data Validation and Cleansing

La quinta fase del ciclo di vita è dedicata alla validazione e alla pulizia dei dati. Il suo obiettivo è garantire che i dati su cui si baseranno le analisi siano accurati, completi, coerenti e privi di anomalie che potrebbero compromettere la validità dei risultati [Erl *et al.*, 2016].

Le attività principali di questa fase comprendono:

- la rilevazione e la gestione dei valori mancanti, che possono essere eliminati, imputati con tecniche statistiche o trattati mediante strategie specifiche del dominio applicativo;
- l'identificazione e il trattamento dei valori anomali, che possono rappresentare errori di misurazione oppure osservazioni legittime ma estreme;
- la verifica della coerenza interna dei dati, ossia il controllo che i valori registrati rispettino i vincoli logici e i *range* attesi per ciascuna variabile;
- la standardizzazione dei formati, delle unità di misura e delle codifiche.

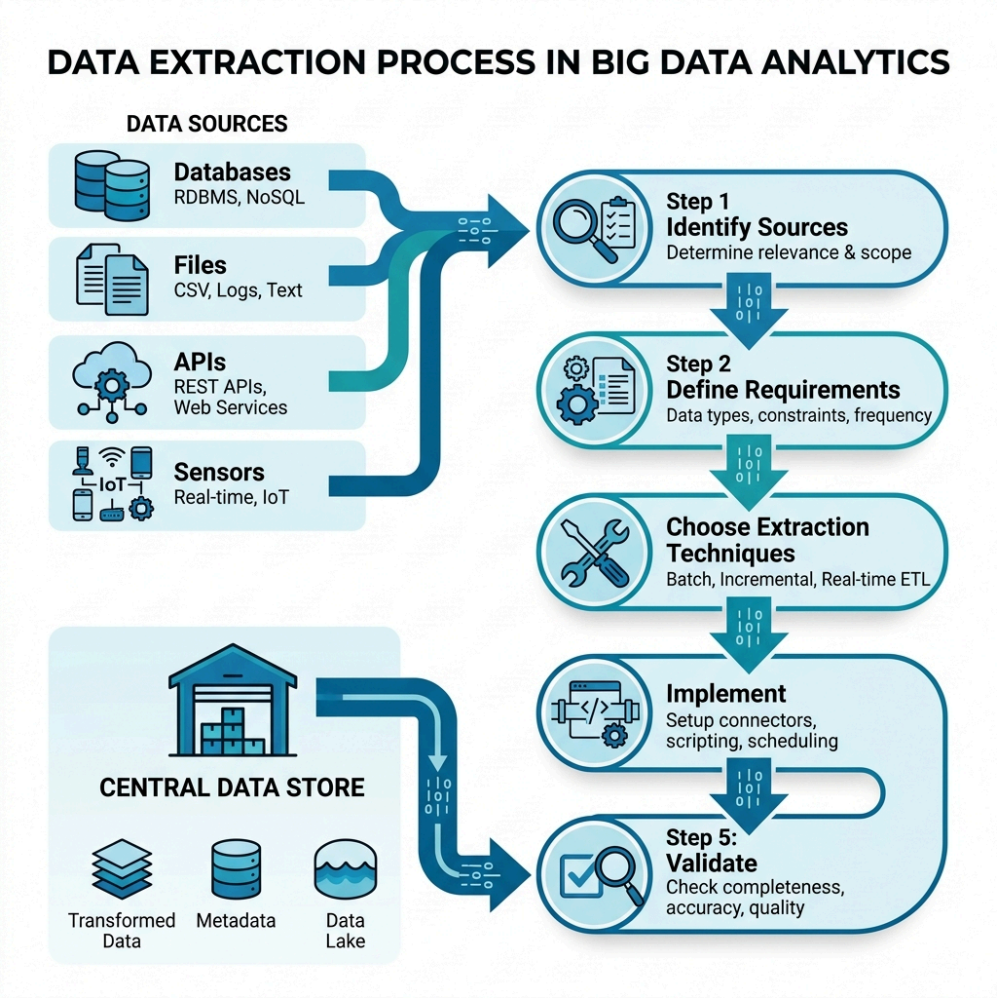


Figura 1.5: Esempio schematico del processo di estrazione dei dati da fonti eterogenee

In ambito clinico, la pulizia dei dati assume un'importanza ancora maggiore rispetto ad altri contesti applicativi. Un errore in un valore di potenza spettrale o un'attribuzione diagnostica errata possono condurre a conclusioni fuorvianti, con potenziali implicazioni sulla comprensione dei fenomeni neurofisiologici studiati.

1.6.6 Data Aggregation and Representation

La sesta fase riguarda l'aggregazione e la rappresentazione dei dati. Una volta che i dati sono stati puliti e validati, è spesso necessario integrarli, combinarli e ristrutturarli per creare una visione unificata e coerente del fenomeno in esame [Erl *et al.*, 2016].

L'aggregazione può operare a diversi livelli. A livello tecnico, essa comprende il collegamento di tabelle provenienti da fonti diverse, la creazione di chiavi di relazione e la definizione di gerarchie dimensionali (Figura 1.6). A livello concettuale, essa include la costruzione di metriche sintetiche che condensano informazioni distribuite su più variabili in indicatori compatti e significativi.

1.6.7 Data Analysis

La settima fase, quella dell'analisi vera e propria, costituisce il cuore operativo del ciclo di vita della *Big Data Analytics*. Durante questa fase, i dati aggregati e rappresentati vengono

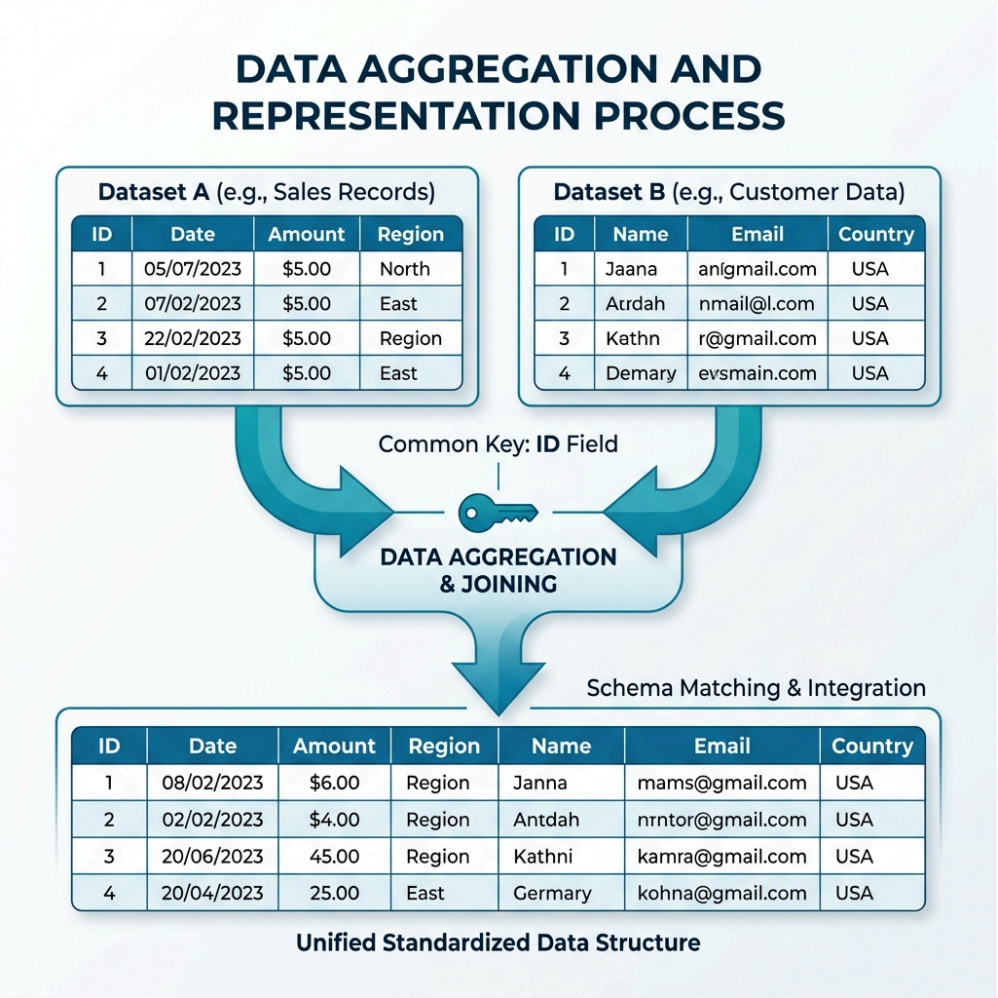


Figura 1.6: Esempio di aggregazione di dati: due dataset vengono combinati utilizzando un campo chiave comune per creare una struttura dati unificata

sottoposti a tecniche analitiche con l'obiettivo di estrarre *pattern*, relazioni, tendenze e anomalie che rispondano alle domande formulate nella fase di valutazione del caso di studio [Erl *et al.*, 2016].

L'analisi dei dati è tipicamente un processo iterativo. L'analista formula un'ipotesi, la verifica attraverso tecniche esplorative o inferenziali, interpreta i risultati e, sulla base di quanto emerge, raffina le ipotesi o ne formula di nuove. Questo ciclo si ripete fino a quando non si raggiunge un livello di comprensione soddisfacente del fenomeno indagato.

Le attività principali di questa fase comprendono:

- l'analisi esplorativa dei dati, che utilizza visualizzazioni e statistiche descrittive per acquisire una comprensione iniziale della struttura e della distribuzione dei dati;
- l'ingegneria delle *feature*, che trasforma le variabili originali in rappresentazioni più informative per le analisi successive;
- l'applicazione di tecniche statistiche inferenziali, quali test di ipotesi, analisi della varianza e modelli di regressione;
- l'applicazione di algoritmi di *machine learning*, sia supervisionati sia non supervisionati, per la classificazione, il *clustering* e la previsione;

- il riconoscimento di *pattern*, che individua regolarità ricorrenti nei dati.

1.6.8 Data Visualization

L'ottava fase del ciclo di vita è dedicata alla visualizzazione dei dati. Il suo obiettivo è comunicare i risultati dell'analisi in modo efficace, chiaro e accessibile a un pubblico che può includere tanto gli analisti tecnici quanto i decisori non specialisti [Erl *et al.*, 2016].

La visualizzazione dei dati non è un mero esercizio estetico, ma una disciplina che richiede competenze specifiche nella scelta delle rappresentazioni grafiche più appropriate per ciascun tipo di informazione. Un grafico mal progettato può occultare *pattern* significativi o, peggio, suggerire conclusioni errate. Al contrario, una visualizzazione ben concepita rende immediatamente percepibili relazioni e tendenze che rimarrebbero nascoste in una tabella numerica. Le principali tipologie di visualizzazione, classificate per finalità, sono riassunte nella Figura 1.7.

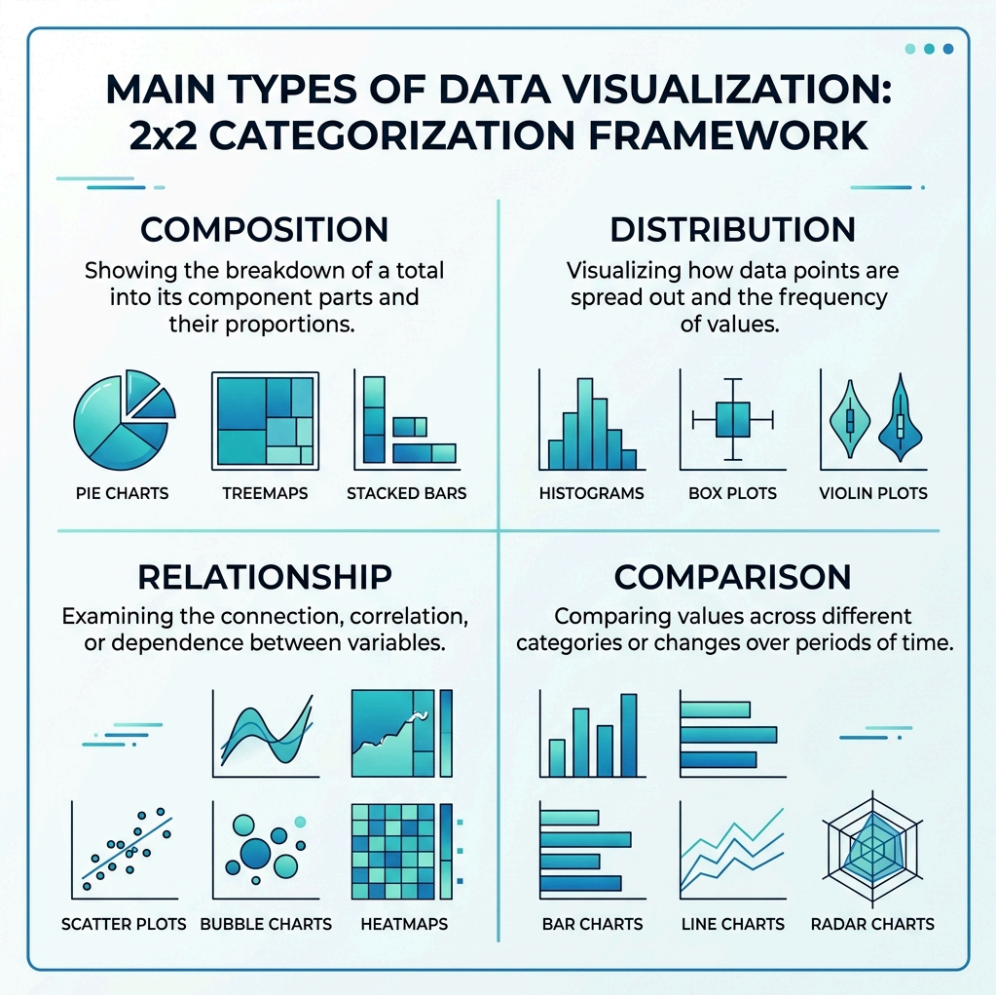


Figura 1.7: Principali tipologie di visualizzazione dei dati classificate per tipo di informazione rappresentata: composizione, distribuzione, relazione e confronto

Le *dashboard* interattive rappresentano lo strumento di visualizzazione più completo, poiché combinano molteplici grafici in un ambiente unificato e consentono all'utente di esplorare i dati attraverso filtri, selezioni e operazioni di *drill-down*. Tra le piattaforme più diffuse per la creazione di visualizzazioni analitiche avanzate figurano strumenti di *Business*

Intelligence come Tableau, Qlik Sense e Microsoft Power BI, che consentono di costruire ambienti interattivi accessibili anche a utenti non specialisti nell'analisi dei dati.

1.6.9 Utilization of Analysis Results

L'ultima fase del ciclo di vita riguarda l'utilizzo dei risultati dell'analisi (*Utilization of Analysis Results*) (Figura 1.7). Questa fase chiude il cerchio aperto nella valutazione del caso di studio, trasformando le conoscenze estratte dai dati in azioni concrete, decisioni operative o nuove linee di indagine [Erl *et al.*, 2016].

L'utilizzo dei risultati può assumere forme diverse a seconda del contesto applicativo:

- in ambito aziendale, i risultati dell'analisi possono orientare strategie commerciali, ottimizzare processi operativi o guidare investimenti;
- in ambito scientifico, essi possono confermare o confutare ipotesi di ricerca, generare nuove domande e contribuire all'avanzamento della conoscenza nel dominio di studio;
- in ambito clinico, i risultati possono supportare la diagnosi, informare la scelta terapeutica o identificare sottopopolazioni di pazienti con caratteristiche distintive.

Questa fase include anche la documentazione dei risultati, la loro comunicazione agli *stakeholder* e la valutazione dell'impatto che l'analisi ha avuto rispetto agli obiettivi inizialmente definiti. È, inoltre, durante questa fase che si identificano i limiti del lavoro svolto e si delineano le possibili direzioni di sviluppo futuro.

Il ciclo di vita che abbiamo descritto in questa sezione non si conclude in modo definitivo con l'utilizzo dei risultati. I nuovi interrogativi che emergono dall'analisi, le limitazioni identificate e le opportunità di approfondimento possono innescare un nuovo ciclo, nel quale le fasi vengono ripercorse con dati aggiuntivi, con tecniche più sofisticate o con domande più mirate. Questa natura ciclica e iterativa è una caratteristica fondamentale della *Big Data Analytics* e rappresenta uno dei motivi per cui essa continua a generare valore nel tempo.

CAPITOLO 2

Introduzione a Microsoft Power BI

In questo capitolo viene presentato Microsoft Power BI, lo strumento di Business Intelligence utilizzato per la realizzazione delle analisi descritte nei capitoli successivi. Dopo una panoramica sulla piattaforma e sul suo posizionamento nel mercato della Business Intelligence, si descrivono l'architettura e i componenti principali dell'ecosistema Power BI. Si illustra quindi il modello dei dati adottato dalla piattaforma, con particolare attenzione allo schema a stella e alla gestione delle relazioni tra tabelle. Il capitolo prosegue con una rassegna delle principali tipologie di visualizzazione disponibili e si conclude con la presentazione dei due linguaggi integrati – M e DAX – che costituiscono il cuore delle operazioni di trasformazione e analisi dei dati.

2.1 Che cos'è Power BI

2.1.1 Definizione e panoramica della piattaforma

Microsoft Power BI è una piattaforma di *Business Intelligence* sviluppata da Microsoft e rilasciata nella sua prima versione commerciale nel luglio 2015. La piattaforma consente di connettere fonti di dati eterogenee, trasformare e modellare i dati, creare visualizzazioni interattive e condividere i risultati delle analisi all'interno di un'organizzazione. Power BI si rivolge a un'ampia platea di utenti, dagli analisti aziendali ai professionisti dell'informatica, fino ai decisori che necessitano di strumenti di consultazione rapida e intuitiva [Russo e Ferrari, 2019].

Le origini della piattaforma risalgono, tuttavia, a un periodo precedente. A partire dal 2010, Microsoft ha sviluppato una serie di componenti aggiuntivi per Excel – tra cui Power Pivot, Power Query e Power View – destinati a potenziare le capacità analitiche del foglio di calcolo. Tali strumenti, inizialmente concepiti come estensioni indipendenti, sono stati progressivamente integrati in un'unica soluzione autonoma, che ha preso il nome di Power BI. Questa evoluzione ha segnato il passaggio da un approccio alla *Business Intelligence* centrato su Excel a una piattaforma dedicata, capace di gestire volumi di dati significativamente maggiori e di offrire funzionalità di condivisione e collaborazione assenti nel foglio di calcolo tradizionale [Microsoft Corporation, 2024].

2.1.2 Posizionamento nel mercato della Business Intelligence

Il mercato degli strumenti di *Business Intelligence* è caratterizzato dalla presenza di numerose piattaforme, ciascuna con peculiarità proprie. Tra le più diffuse, oltre a Power BI, figurano Tableau, sviluppato da Salesforce, e Qlik Sense, prodotto da Qlik Technologies. Per orientarsi in un panorama così articolato, uno dei riferimenti più autorevoli è il *Magic Quadrant* pubblicato annualmente dalla società di ricerca Gartner, che classifica i principali fornitori di soluzioni analitiche e di *Business Intelligence* secondo due dimensioni: la completezza della visione strategica e la capacità di esecuzione.

Nella Figura 2.1 riportiamo il posizionamento delle principali piattaforme nel *Magic Quadrant* di Gartner per l'anno 2026. Come è possibile osservare, Microsoft Power BI si colloca stabilmente nel quadrante dei *Leader* – la posizione riservata ai fornitori che combinano una visione strategica ampia con una solida capacità di realizzarla – confermando un primato mantenuto per diversi anni consecutivi [Research, 2026].

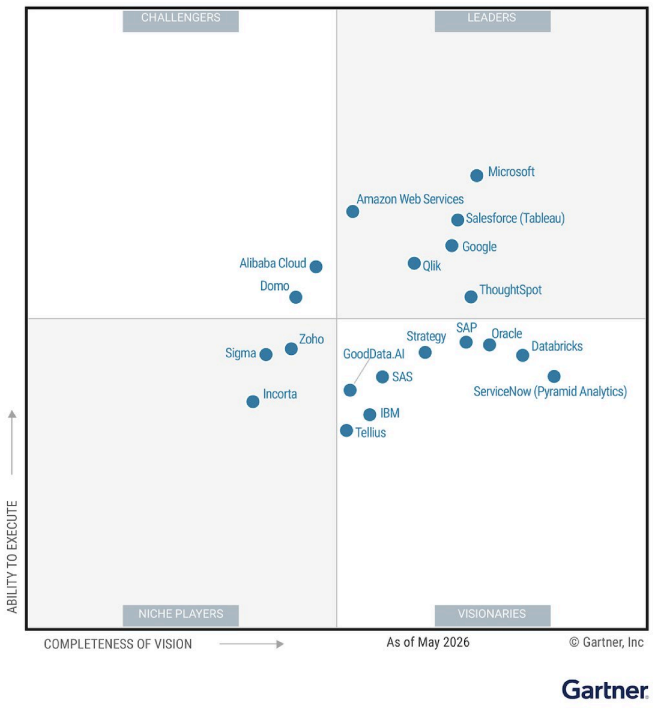


Figura 2.1: Posizionamento delle principali piattaforme di *Business Intelligence* nel *Magic Quadrant* di Gartner (2026). Microsoft Power BI si colloca nel quadrante dei *Leader*

2.1.3 Confronto con altre piattaforme

Per comprendere le ragioni del successo di Power BI, è utile un confronto sintetico con le due principali alternative presenti sul mercato: Tableau e Qlik Sense.

Tableau è riconosciuto per l'eccellenza nelle visualizzazioni e per la flessibilità nell'esplorazione interattiva dei dati. Il suo punto di forza risiede nella capacità di produrre grafici sofisticati con un'interfaccia altamente intuitiva, basata sul paradigma del *drag-and-drop*. Tableau risulta particolarmente adatto per contesti in cui la qualità estetica e la profondità delle visualizzazioni costituiscono requisiti prioritari. Tuttavia, il costo delle licenze è generalmente più elevato rispetto a Power BI, e l'integrazione con l'ecosistema Microsoft risulta meno diretta [Russo e Ferrari, 2019].

Qlik Sense si distingue per il suo motore associativo, che consente di esplorare i dati senza vincoli di percorso predefiniti, permettendo all'utente di navigare liberamente tra le dimensioni del *dataset*. Questa architettura è particolarmente efficace per l'analisi esplorativa e per l'identificazione di relazioni non previste. Per contro, la curva di apprendimento è generalmente più ripida e la comunità di utenti è meno estesa rispetto a quelle di Power BI e Tableau.

Power BI si contraddistingue per la profonda integrazione con l'ecosistema Microsoft – in particolare con Excel, Azure, SharePoint e Teams – per il rapporto favorevole tra funzionalità offerte e costo della licenza, e per la vasta comunità di utenti e sviluppatori che alimenta un ricco ecosistema di risorse formative, componenti aggiuntivi e visualizzazioni personalizzate. In ragione di tali caratteristiche, Power BI rappresenta una scelta particolarmente adatta per le organizzazioni che operano già all'interno dell'infrastruttura Microsoft e che necessitano di uno strumento capace di coprire l'intero ciclo analitico, dalla connessione ai dati alla condivisione dei risultati.

2.2 Architettura e componenti principali

L'ecosistema Power BI si articola in tre componenti principali, ciascuno progettato per rispondere a esigenze specifiche del processo analitico. Nella Figura 2.2 illustriamo schematicamente l'architettura complessiva della piattaforma, evidenziando le relazioni funzionali tra i tre componenti.

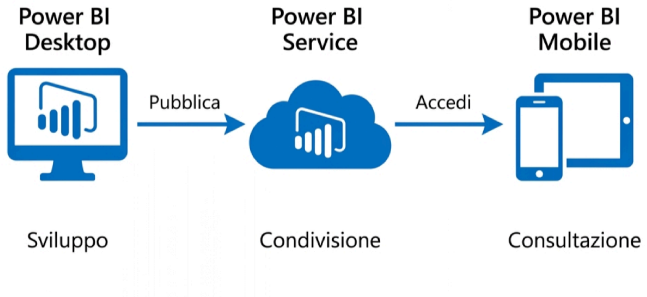


Figura 2.2: Architettura dell'ecosistema Power BI: i tre componenti principali – Desktop, Service e Mobile – e le loro interazioni

2.2.1 Power BI Desktop

Power BI Desktop è l'applicazione gratuita installabile su sistemi Windows che costituisce l'ambiente di sviluppo principale della piattaforma. In Power BI Desktop l'utente svolge tutte le attività fondamentali del processo analitico: la connessione alle sorgenti dei dati, la trasformazione e la pulizia dei dati attraverso Power Query, la costruzione del modello relazionale, la definizione di misure e colonne calcolate in DAX, e la creazione di visualizzazioni interattive organizzate in pagine di *report* [Microsoft Corporation, 2024].

L'interfaccia di Power BI Desktop è organizzata intorno a tre viste principali, accessibili tramite le icone posizionate nella barra laterale sinistra:

- la vista *Report*, che rappresenta l'ambiente di progettazione delle visualizzazioni;
- la vista *Dati*, che consente di esplorare il contenuto delle tabelle caricate nel modello;

- la vista *Modello*, che permette di visualizzare e gestire le relazioni tra le tabelle.

Power BI Desktop supporta la connessione a un numero molto ampio di sorgenti di dati. Tra le principali figurano i file locali (Excel, CSV, JSON, XML), i *database* relazionali (SQL Server, Oracle, MySQL, PostgreSQL), i servizi *cloud* (Azure SQL Database, Google BigQuery, Amazon Redshift), le interfacce di programmazione (API) e i flussi di dati in tempo reale. Questa versatilità nella connessione alle sorgenti rende Power BI adatto a contesti applicativi molto diversi tra loro.

2.2.2 Power BI Service

Power BI Service è la componente *cloud* della piattaforma, accessibile tramite browser all'indirizzo `app.powerbi.com`. Il suo ruolo principale consiste nel consentire la pubblicazione, la condivisione e la fruizione collaborativa dei *report* e delle *dashboard* creati in Power BI Desktop [Russo e Ferrari, 2019].

Una volta completato lo sviluppo in Power BI Desktop, l'utente può pubblicare il proprio lavoro nel servizio *cloud*, dove diventa accessibile agli altri membri dell'organizzazione in base ai permessi configurati. Power BI Service offre, inoltre, funzionalità aggiuntive rispetto alla versione Desktop; questi sono:

- la creazione di *dashboard*, ovvero pagine sintetiche composte da elementi provenienti da *report* diversi, progettate per offrire una panoramica immediata delle informazioni più rilevanti;
- la pianificazione dell'aggiornamento automatico dei dati, che consente di mantenere i *report* costantemente allineati con le sorgenti originali;
- la configurazione di avvisi, che notificano gli utenti quando una metrica supera una soglia predefinita;
- la gestione della sicurezza a livello di riga, che permette di controllare quali dati ciascun utente può visualizzare.

2.2.3 Power BI Mobile

Power BI Mobile è l'applicazione per dispositivi mobili (iOS, Android e Windows) che consente di consultare *report* e *dashboard* pubblicati nel servizio *cloud*. L'applicazione è progettata per offrire un'esperienza di fruizione ottimizzata per schermi di dimensioni ridotte, mantenendo le funzionalità di interazione con i dati – filtri, selezioni, operazioni di *drill-down* – in un formato adattato al contesto mobile.

Power BI Mobile non è uno strumento di sviluppo, ma di consultazione; il suo valore risiede nella possibilità di accedere ai risultati delle analisi in qualsiasi momento e luogo, una caratteristica particolarmente rilevante per i decisori che necessitano di informazioni aggiornate senza essere vincolati a una postazione fissa.

2.3 Il modello dei dati in Power BI

2.3.1 Il concetto di modello relazionale

Il modello dei dati rappresenta la struttura logica sulla quale si fondano tutte le analisi condotte in Power BI. Un modello dei dati ben progettato costituisce il presupposto essenziale

per la correttezza dei calcoli, l'efficienza delle interrogazioni e la chiarezza delle visualizzazioni. In assenza di un modello solido, anche le misure più sofisticate e le visualizzazioni più curate rischiano di produrre risultati errati o fuorvianti.

In Power BI, il modello dei dati si basa sul paradigma relazionale; le informazioni sono organizzate in tabelle distinte, collegate tra loro attraverso relazioni definite su colonne condivise. Questo approccio consente di evitare la ridondanza dei dati, di mantenere la coerenza delle informazioni e di sfruttare le relazioni per aggregare e filtrare i dati in modo dinamico.

2.3.2 Lo schema a stella

La struttura più diffusa e raccomandata per i modelli di dati in Power BI è lo *schema a stella*, un modello di organizzazione dei dati proposto originariamente da Ralph Kimball nell'ambito del *data warehousing* [Kimball e Ross, 2013]. Nella Figura 2.3 illustriamo la struttura tipica di uno schema a stella, che prende il nome dalla disposizione grafica che le tabelle assumono quando le relazioni vengono rappresentate visivamente.

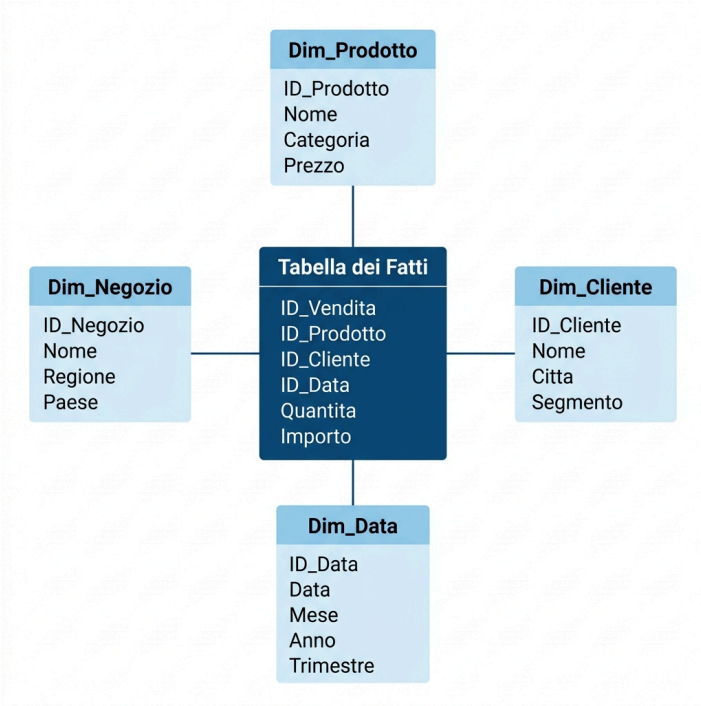


Figura 2.3: Rappresentazione schematica dello schema a stella: una tabella dei fatti centrale collegata a più tabelle dimensionali

Lo schema a stella si compone di due tipologie di tabelle:

- Le *tabelle dei fatti*, che contengono i dati quantitativi e le misure oggetto dell'analisi. Ogni riga di una tabella dei fatti rappresenta un evento o una transazione – ad esempio, una vendita, una visita clinica, una misurazione strumentale. Le colonne di queste tabelle comprendono le metriche numeriche (importi, quantità, valori misurati) e le chiavi esterne che collegano ciascun evento alle tabelle dimensionali pertinenti.
- Le *tabelle delle dimensioni*, che contengono gli attributi descrittivi e le categorie utilizzate per contestualizzare, filtrare e raggruppare i dati contenuti nelle tabelle dei fatti. Esempi tipici di tabelle delle dimensioni sono quelle relative alle date, ai prodotti, ai clienti, alle aree geografiche o, in ambito clinico, ai pazienti e alle diagnosi.

Lo schema a stella favorisce la semplicità del modello e l'efficienza delle interrogazioni: le tabelle delle dimensioni fungono da filtri che propagano le selezioni dell'utente verso la tabella dei fatti, consentendo di ottenere aggregazioni dinamiche senza la necessità di scrivere interrogazioni complesse.

2.3.3 Gestione delle relazioni

Le relazioni tra tabelle costituiscono l'elemento strutturale che consente al motore di Power BI di propagare i filtri e di calcolare le misure aggregate in modo corretto. Nella vista *Modello* di Power BI Desktop le relazioni sono rappresentate graficamente come linee che collegano le colonne condivise tra due tabelle.

Ogni relazione è caratterizzata da due proprietà fondamentali:

- La *cardinalità*, che specifica il tipo di corrispondenza tra le righe delle due tabelle coinvolte. Le cardinalità supportate da Power BI sono: uno-a-molti (la più comune nello schema a stella), uno-a-uno, molti-a-molti e molti-a-uno.
- La *direzione del filtro incrociato*, che determina il verso in cui i filtri applicati a una tabella si propagano alle tabelle correlate. In una relazione uno-a-molti, il filtro si propaga tipicamente dalla tabella delle dimensioni (lato "uno") alla tabella dei fatti (lato "molti"); tuttavia, è possibile configurare una propagazione bidirezionale quando necessario.

Una corretta definizione delle relazioni e della loro direzione di filtro è fondamentale per garantire che le misure calcolate in DAX producano risultati coerenti con le attese. Relazioni mal configurate possono portare a conteggi duplicati, aggregazioni errate o filtri che non si propagano come previsto.

2.4 Le visualizzazioni in Power BI

2.4.1 Tipologie di oggetti visivi

Power BI mette a disposizione un ampio catalogo di oggetti visivi nativi, ciascuno progettato per rappresentare efficacemente una specifica tipologia di informazione. La scelta dell'oggetto visivo più appropriato dipende dalla natura dei dati e dalla domanda analitica a cui si intende rispondere. Nella Figura 2.4 presentiamo una panoramica delle principali tipologie di visualizzazione disponibili in Power BI, organizzate per categoria funzionale.

Tra le categorie più utilizzate figurano:

- *i grafici a barre e a colonne*, adatti al confronto tra valori categoriali;
- *i grafici a linee*, indicati per rappresentare l'andamento di una variabile nel tempo;
- *i grafici a torta e ad anello*, utilizzati per mostrare la composizione percentuale di un insieme;
- *le tabelle e le matrici*, che consentono di visualizzare i dati in formato tabellare con possibilità di raggruppamento e sub-totali;
- *le mappe geografiche*, utili per rappresentare la distribuzione spaziale dei dati;
- *gli indicatori KPI e le card*, che sintetizzano una singola metrica in un valore numerico immediatamente leggibile;
- *i grafici a dispersione*, che evidenziano la relazione tra due variabili numeriche.

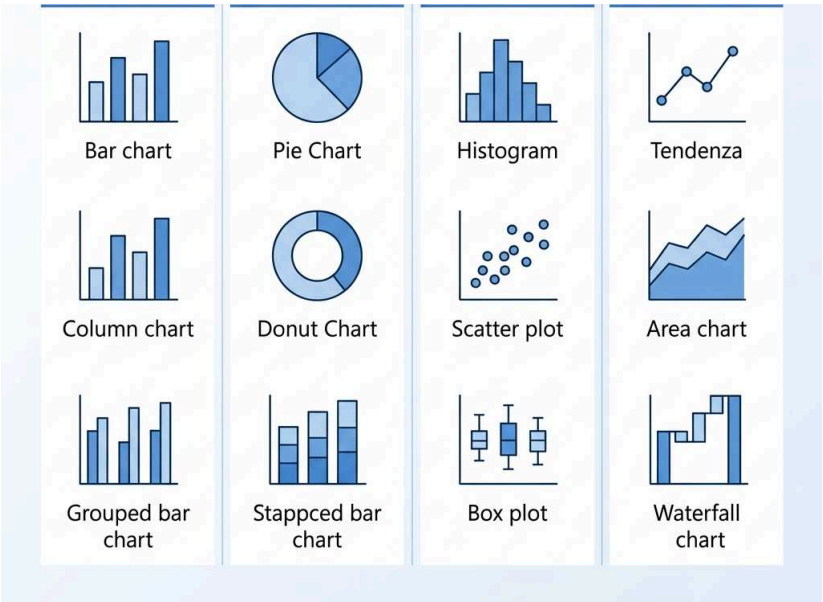


Figura 2.4: Panoramica delle principali tipologie di oggetti visivi disponibili in Power BI, organizzate per categoria funzionale

Oltre agli oggetti visivi nativi, Power BI consente di importare visualizzazioni personalizzate sviluppate dalla comunità e disponibili nel *marketplace* di Microsoft, ampliando ulteriormente le possibilità di rappresentazione dei dati.

2.4.2 Report e dashboard

In Power BI, i termini *report* e *dashboard* designano due concetti distinti, sebbene spesso utilizzati in modo intercambiabile nel linguaggio comune.

Un *report* è un documento analitico composto da una o più pagine, ciascuna contenente un insieme di oggetti visivi connessi a un unico *dataset*. Il *report* rappresenta lo strumento principale per l'esplorazione dettagliata dei dati e viene creato in Power BI Desktop. Ogni pagina di un *report* può contenere grafici, tabelle, filtri e altri elementi interattivi che consentono all'utente di analizzare i dati da prospettive diverse.

Una *dashboard*, al contrario, è una singola pagina creata nel servizio *cloud* che può aggregare elementi visivi provenienti da *report* diversi e da *dataset* differenti. La *dashboard* è progettata per offrire una sintesi immediata delle informazioni più rilevanti, fungendo da punto di ingresso per l'esplorazione approfondita: cliccando su un elemento della *dashboard*, l'utente viene indirizzato al *report* di origine per un'analisi di dettaglio.

2.4.3 Filtri, slicer e interattività

Uno degli aspetti distintivi di Power BI è il livello di interattività che gli oggetti visivi offrono all'utente. L'interattività in Power BI si manifesta attraverso diversi meccanismi:

- il *cross-filtering*, per cui la selezione di un elemento in un oggetto visivo filtra automaticamente tutti gli altri oggetti visivi della stessa pagina, evidenziando le relazioni tra le diverse dimensioni dei dati;
- gli *slicer*, ovvero oggetti visivi dedicati al filtraggio, che consentono all'utente di selezionare valori specifici di una dimensione e di osservare come le metriche variano in funzione della selezione effettuata;

- il riquadro *Filtri*, che permette di applicare filtri a tre livelli distinti: sull'intero *report*, sulla singola pagina o sul singolo oggetto visivo;
- il *drill-down* e il *drill-up*, che consentono di navigare all'interno di gerarchie dimensionali, passando da un livello di aggregazione a uno di maggiore dettaglio e viceversa.

Questi meccanismi di interattività trasformano un *report* statico in uno strumento di esplorazione dinamica, in cui l'utente può formulare domande ai dati e ottenere risposte immediate attraverso la manipolazione diretta delle visualizzazioni.

2.5 I linguaggi integrati: M e DAX

Power BI integra due linguaggi specializzati per le operazioni di trasformazione e analisi dei dati: il linguaggio M, utilizzato in Power Query per la fase di accesso e preparazione dei dati, e il linguaggio DAX, impiegato per la creazione di misure e colonne calcolate nel modello dei dati [Russo e Ferrari, 2019].

2.5.1 Il linguaggio M e Power Query

Power Query è il componente di Power BI dedicato alla connessione, all'importazione e alla trasformazione dei dati provenienti dalle sorgenti esterne. L'interfaccia grafica di Power Query – denominata *Editor di Power Query* – consente di definire sequenze di trasformazioni in modo visivo, attraverso un sistema di passi registrati che documentano ogni operazione applicata ai dati.

Dietro l'interfaccia grafica, ogni trasformazione viene tradotta automaticamente in codice scritto nel linguaggio M, un linguaggio funzionale progettato specificamente per le operazioni di *data shaping*. L'utente può visualizzare e modificare direttamente il codice M attraverso l'Editor avanzato, una funzionalità utile quando le trasformazioni richieste superano le possibilità dell'interfaccia grafica.

Le operazioni più comuni eseguite in Power Query comprendono:

- la rimozione di colonne e righe non necessarie;
- la modifica dei tipi di dati;
- la sostituzione e la gestione dei valori mancanti o errati;
- l'unione e l'accodamento di tabelle provenienti da fonti diverse;
- la creazione di colonne derivate mediante espressioni condizionali;
- la trasposizione, il *pivot* e l'*unpivot* delle tabelle.

La separazione tra la fase di trasformazione (gestita da Power Query e dal linguaggio M) e la fase di analisi (gestita dal modello dei dati e dal linguaggio DAX) costituisce un principio architetturale fondamentale di Power BI: i dati vengono puliti e strutturati *prima* di entrare nel modello, garantendo che le analisi successive operino su una base informativa coerente e affidabile.

2.5.2 Il linguaggio DAX

Il linguaggio DAX è il linguaggio di espressione utilizzato in Power BI per la definizione di calcoli all'interno del modello dei dati. DAX è un linguaggio funzionale, la cui sintassi richiama per alcuni aspetti quella delle formule di Excel, ma che si distingue per la capacità di operare su insiemi di righe e di gestire il contesto di valutazione in modo dinamico [Russo e Ferrari, 2019].

In Power BI, il linguaggio DAX viene impiegato per la creazione di due tipologie principali di oggetti calcolati:

- Le *colonne calcolate*, che aggiungono una nuova colonna a una tabella esistente. Il valore di una colonna calcolata viene determinato riga per riga al momento del caricamento dei dati e viene memorizzato fisicamente nel modello. Le colonne calcolate sono utili quando il risultato del calcolo deve essere utilizzato come criterio di filtro, raggruppamento o ordinamento.
- Le *misure*, che definiscono calcoli valutati dinamicamente in risposta al contesto di filtro corrente. A differenza delle colonne calcolate, le misure non sono associate a una singola riga, ma producono un risultato aggregato che varia in funzione delle selezioni effettuate dall'utente nelle visualizzazioni. Le misure rappresentano lo strumento fondamentale per la creazione di KPI e metriche analitiche.

2.5.3 Il concetto di contesto in DAX

Il concetto di contesto costituisce il fondamento del funzionamento del linguaggio DAX e rappresenta l'aspetto che più lo differenzia dalle formule tradizionali di Excel. Il contesto determina *quali righe* vengono considerate nel calcolo di un'espressione DAX e si declina in due forme distinte [Russo e Ferrari, 2019], ovvero:

- Il *contesto di riga*, che identifica la singola riga corrente durante la valutazione di un'espressione. Il contesto di riga si attiva tipicamente nelle colonne calcolate e nelle funzioni iterative, come `SUMX`, `AVERAGEX` e `FILTER`. Quando un'espressione opera in contesto di riga, essa ha accesso ai valori della riga corrente e può utilizzarli per calcoli specifici.
- Il *contesto di filtro*, che definisce l'insieme delle righe "visibili" a un'espressione in un dato momento. Il contesto di filtro è determinato dalle selezioni attive nel *report* – filtri, *slicer*, selezioni sugli oggetti visivi – e dalle relazioni definite nel modello dei dati. Quando l'utente seleziona un valore in un filtro, il contesto di filtro si restringe di conseguenza, e tutte le misure vengono ricalcolate considerando soltanto le righe che soddisfano i criteri di filtraggio.

La comprensione del contesto è essenziale per scrivere espressioni DAX corrette. Funzioni come `CALCULATE`, `ALL`, `FILTER` e `RELATED` operano modificando o navigando il contesto, consentendo di creare calcoli sofisticati, come percentuali sul totale, medie mobili, confronti temporali e aggregazioni condizionate.

La padronanza del linguaggio DAX e del concetto di contesto costituisce una competenza fondamentale per chiunque utilizzi Power BI in modo avanzato; infatti, la capacità di costruire misure corrette e significative determina in larga misura la qualità e l'affidabilità delle analisi condotte attraverso la piattaforma.

CAPITOLO 3

Descrizione dei dati di partenza e attività di ETL

Il presente capitolo si pone l'obiettivo di illustrare dettagliatamente il processo di acquisizione, preparazione e modellazione dei dati clinici che costituiscono il fondamento del nostro studio. Nella prima parte verranno fornite le basi neurofisiologiche necessarie per comprendere i dati analizzati, esplorando i principi dell'elettroencefalografia e il significato clinico delle potenze spettrali. Successivamente, verrà presentato in dettaglio il dataset clinico di partenza, analizzandone la struttura e le distribuzioni patologiche. Infine, sarà descritto rigorosamente l'intero processo di ETL (Extraction, Transformation, Loading) e Data Engineering, giustificando le scelte metodologiche implementate sia nell'ambiente di trasformazione visiva di Power Query, sia mediante il linguaggio analitico DAX, allo scopo di predisporre un modello semantico robusto e affidabile per l'interrogazione tramite dashboard.

3.1 L'elettroencefalogramma: cenni sui dati

L'elettroencefalografia rappresenta una delle metodiche non invasive più diffuse ed efficaci per l'esplorazione funzionale del cervello umano. Essa si basa sulla misurazione e sulla registrazione dell'attività elettrica cerebrale, generata prevalentemente dai potenziali post-sinaptici dei neuroni della corteccia. L'elevata risoluzione temporale di questo esame permette di catturare variazioni neurofisiologiche nell'ordine dei millisecondi, il che lo rende uno strumento fondamentale sia nella diagnostica neurologica (ad esempio per l'epilessia o le encefalopatie), sia nella ricerca psichiatrica per lo studio delle alterazioni funzionali legate ai disturbi dell'umore, all'ansia e alla schizofrenia.

Per garantire la standardizzazione e la riproducibilità delle misurazioni a livello globale, il posizionamento degli elettrodi sullo scalpo del paziente segue rigorosamente il sistema internazionale 10-20 [Prichep e John, 1992]. Tale convenzione si basa sulla misurazione di distanze proporzionali (il 10% o il 20% della distanza totale) tra specifici punti di repere anatomici del cranio, come il nasion e l'inion. Nella Figura 3.1 illustriamo lo schema concettuale di questo sistema di posizionamento.

Nel contesto del nostro lavoro, facciamo riferimento a un set standard di diciannove elettrodi, distribuiti strategicamente per mappare l'intera superficie corticale: prefrontali (Fp1, Fp2), frontali (F3, F4, F7, F8, Fz), centrali (C3, C4, Cz), parietali (P3, P4, Pz), temporali (T3, T4, T5, T6) e occipitali (O1, O2). Questa configurazione permette di analizzare la topografia dell'attività elettrica e di individuare eventuali asimmetrie o anomalie focali.

L'analisi del segnale EEG si basa sulla sua scomposizione nel dominio della frequenza, tipicamente ottenuta tramite l'applicazione della Trasformata Rapida di Fourier. Questo

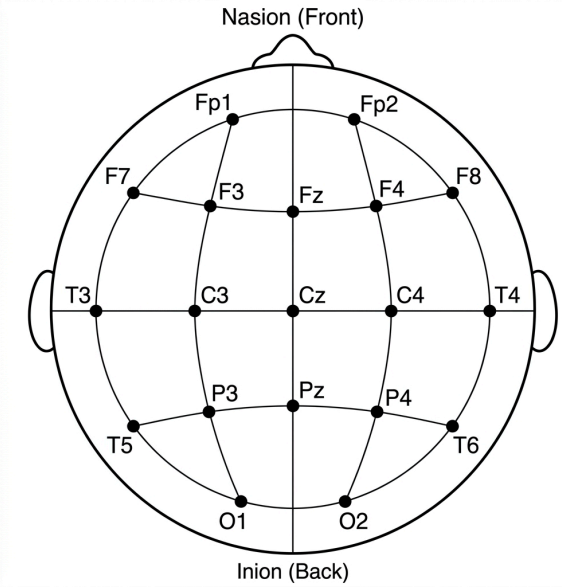


Figura 3.1: Rappresentazione schematica del sistema internazionale 10-20 per il posizionamento degli elettrodi EEG.

processo permette di isolare e quantificare l'energia del segnale all'interno di specifici range di frequenza, comunemente denominati bande. Ciascuna di esse è fisiologicamente associata a determinati stati cognitivi e comportamentali. In particolare, si riconoscono le seguenti bande:

- *Banda Delta* (0,5–4 Hz): rappresenta le oscillazioni più lente ed è predominante durante le fasi di sonno profondo. La sua presenza marcata in un soggetto in stato di veglia costituisce un forte indicatore di lesione cerebrale o grave disfunzione corticale.
- *Banda Theta* (4–8 Hz): è tipicamente associata a stati di sonnolenza, transizione verso il sonno, processi di meditazione profonda e consolidamento mnestico. Un suo eccesso durante le attività che richiedono attenzione può suggerire deficit di concentrazione o affaticamento mentale.
- *Banda Alpha* (8–13 Hz): costituisce il ritmo dominante in un soggetto adulto sano in stato di rilassamento vigile ad occhi chiusi, con massima espressione nelle regioni posteriori e occipitali. L'attenuazione o l'assenza di tale ritmo può indicare stati di forte ansia o attivazione corticale anomala.
- *Banda Beta* (13–30 Hz): è correlata all'attenzione attiva, alla concentrazione, alla risoluzione di problemi e alle funzioni esecutive superiori, localizzandosi prevalentemente nelle aree frontali e prefrontali.
- *Banda Gamma* (>30 Hz): rappresenta l'attività a più alta frequenza ed è intimamente legata ai processi cognitivi superiori, all'integrazione multisensoriale e al cosiddetto *binding* neuronale, ovvero la capacità del cervello di unificare diverse percezioni in un'unica esperienza cosciente.

Una volta separato il segnale in queste bande, è possibile calcolarne la potenza spettrale, definita come l'area sottesa allo spettro di potenza per quel determinato intervallo di frequenza. Essa fornisce una misura quantitativa dell'energia sviluppata dai circuiti neuronali. Attraverso l'Analisi Quantitativa dell'EEG, queste misurazioni vengono elaborate

statisticamente e mappate, trasformando i complessi tracciati temporali in metriche oggettive e numeriche. Questo passaggio è di importanza cruciale per il nostro progetto, poiché costituisce il fondamento matematico su cui opereranno i modelli di *Data Science* e le rappresentazioni grafiche all'interno delle dashboard, permettendoci di indagare i meccanismi neurofisiologici nascosti nelle patologie psichiatriche.

3.2 Il dataset clinico

Il cuore analitico di questo progetto di tesi è costituito da un esteso dataset clinico denominato *EEG.machinelearning_data_BRMH*, contenente i tracciati elettroencefalografici e le relative misurazioni neuropsicologiche di una vasta coorte di pazienti [Prichep e John, 1992]. Nello specifico, il campione comprende 941 soggetti, i quali sono stati sottoposti a registrazioni EEG standardizzate al fine di indagare le basi neurofisiologiche di diverse patologie neurologiche e psichiatriche. La struttura originale del dataset si presenta in formato tabellare ed è caratterizzata da un'elevata dimensionalità. Oltre alle informazioni anagrafiche e diagnostiche, la maggior parte delle variabili quantifica l'energia delle onde cerebrali, restituendo una "fotografia" dell'attività elettrica del paziente al momento dell'esame. Possiamo suddividere i dati presenti in tre macro-categorie logiche:

- *Variabili anagrafiche e demografiche*: includono parametri quali l'età del paziente (`age`), il sesso (`sex`), gli anni di scolarizzazione formale (`education`) e il Quoziente Intellettivo misurato tramite test clinici (`IQ`).
- *Variabili cliniche*: forniscono la classificazione patologica del paziente, suddivisa in una macro-diagnosi (`main.disorder`) e in una diagnosi più specifica (`specific.disorder`).
- *Variabili spettrali*: rappresentano il nucleo matematico dell'esame EEG. Per ciascuno dei diciannove elettrodi del sistema 10-20, il dataset riporta la potenza spettrale calcolata per le cinque bande di frequenza (Delta, Theta, Alpha, Beta, Gamma). Di conseguenza, il tracciato di ogni paziente è descritto da 95 colonne distinte (19 elettrodi $\times$ 5 bande).

A seguire forniamo un riepilogo delle caratteristiche e della nomenclatura tecnica per le colonne di maggiore interesse all'interno dell'analisi:

1. *`no.`*: identificativo univoco del paziente (impostato come valore testuale);
2. *`age`*: età anagrafica del paziente, in formato numerico;
3. *`sex`*: sesso del paziente (M/F);
4. *`education`*: anni di istruzione formale completati;
5. *`IQ`*: quoziente intellettivo globale misurato;
6. *`main.disorder`*: categoria diagnostica principale attribuita al paziente;
7. *`specific.disorder`*: sottotipo diagnostico clinico dettagliato;
8. *`AB.A.delta.a.FP1...`*: misurazioni in formato numerico continuo che rappresentano la potenza assoluta della banda in quello specifico elettrodo;

Uno degli aspetti più rilevanti di questo dataset è la ricchezza e l'eterogeneità delle categorie diagnostiche rappresentate. Il campione non si limita a studiare una singola patologia, ma raccoglie le firme elettriche di molteplici condizioni cliniche, consentendo, così,

analisi comparative e la ricerca di biomarcatori trasversali. In particolare, le macro-diagnosi documentate (`main.disorder`) includono: disturbi dell'umore (come la depressione maggiore e il disturbo bipolare), disturbi d'ansia, schizofrenia, disturbi da dipendenza, disturbo ossessivo-compulsivo (DOC), disturbi correlati a trauma e stress, nonché un gruppo di controllo composto da soggetti sani (*Healthy control*). Infine, è opportuno sottolineare un aspetto fondamentale relativo all'etica della ricerca. Trattandosi di dati biomedici altamente sensibili, tutte le informazioni presenti nel dataset sono state preventivamente e rigorosamente anonimizzate in conformità alle normative sanitarie vigenti sulla privacy. L'identificativo del paziente (`no.`) è l'unico elemento di tracciamento disponibile, garantendo l'impossibilità di risalire all'identità reale dei soggetti coinvolti, pur permettendo il pieno svolgimento delle attività di *Data Science* in completa sicurezza.

3.3 Il processo di ETL: inquadramento concettuale

Prima di procedere con l'analisi esplorativa e diagnostica dei dati, è di vitale importanza garantire che le informazioni grezze fornite dal dataset siano affidabili, strutturalmente corrette e idonee per l'elaborazione informatica avanzata. A tal fine, il flusso di lavoro di analisi dei dati si affida tipicamente a una sequenza standardizzata di operazioni nota come processo di estrazione, trasformazione e caricamento, universalmente indicato con l'acronimo ETL. Tale paradigma metodologico racchiude le tre macro-fasi fondamentali per la corretta preparazione del dato, fungendo da ponte tra le fonti originarie disomogenee e i sistemi di destinazione finale.

Analizzando nel dettaglio l'architettura concettuale di questo processo, possiamo delineare i seguenti passaggi cardine:

- *Estrazione*: rappresenta il prelievo dei dati grezzi dai sistemi sorgente. Tali sistemi possono presentare enormi variabilità in termini di formato, spaziando da semplici file di testo o fogli di calcolo fino a complessi database relazionali e non relazionali. L'obiettivo primario in questa fase è l'acquisizione sicura e completa del dato, senza alterarne in alcun modo la natura originaria.

- *Trasformazione e pulizia*: costituisce indubbiamente la fase più delicata, complessa e onerosa in termini di risorse computazionali e di tempo umano. Durante questo passaggio, il dato viene sottoposto a una rigorosa validazione strutturale e semantica. Ciò include la sanificazione delle informazioni (come la gestione accurata dei valori nulli o anomali), la standardizzazione dei formati tipizzati e, aspetto ancora più strategico, l'ingegnerizzazione di nuove funzionalità matematiche o categoriche, necessarie per arricchire il potenziale informativo della base di dati originale.

- *Caricamento*: si tratta del trasferimento definitivo e strutturato dei dati ormai puliti e arricchiti all'interno del sistema di destinazione. Nel panorama della moderna *Business Intelligence*, tale destinazione è solitamente un modello semantico ottimizzato in memoria o un repository di dati aziendale, pronto per essere interrogato in tempo reale dagli algoritmi di esplorazione e dai motori di visualizzazione.

L'importanza di applicare un rigoroso processo di estrazione, trasformazione e caricamento assume un ruolo critico, quasi irrinunciabile, quando si opera all'interno del delicato contesto biomedico. La natura intrinseca della sperimentazione clinica e della raccolta di segnali neurofisiologici, come nel caso dell'elettroencefalogramma, è fisiologicamente soggetta a notevoli fonti di rumore e instabilità. I dataset clinici sono costantemente affetti da artefatti biologici o strumentali, da errori di trascrizione manuale da parte del personale medico, da valori parzialmente mancanti dovuti all'interruzione dei protocolli sperimentali, oppure

da disallineamenti di formato nei referti. L'immissione di dati non accuratamente sanificati all'interno di un sistema per l'analisi dei dati comporterebbe conseguenze metodologiche disastrose. Essa rischierebbe non solo di produrre distorsioni gravi nelle misurazioni matematiche aggregate (come, ad esempio, nello studio della correlazione tra quoziente intellettivo e potenza spettrale delle singole bande di frequenza), ma condurrebbe inevitabilmente l'analista a trarre conclusioni diagnostiche totalmente errate, inquinando la validità dell'intera ricerca scientifica.

Al fine di implementare concretamente questa delicata infrastruttura metodologica all'interno del presente progetto di tesi, abbiamo adottato Power Query, l'avanzato motore di preparazione dei dati strettamente integrato nell'ecosistema Microsoft. Attraverso la sua robusta interfaccia visiva e la flessibilità del suo linguaggio funzionale di programmazione sottostante, Power Query ci ha fornito gli strumenti ideali per orchestrare l'intero flusso di lavoro. Questo ha permesso di garantire non soltanto l'esattezza metodologica delle singole trasformazioni, ma anche la loro completa automazione, assicurando, così, una piena scalabilità del modello e l'assoluta riproducibilità delle procedure. Nelle sezioni che seguono, analizzeremo nel massimo dettaglio ciascuna delle tre fasi di questo processo, illustrando le soluzioni ingegneristiche adottate per risolvere le criticità del nostro dataset clinico di partenza.

3.4 Fase di estrazione

La fase di estrazione costituisce il primo passaggio operativo della nostra campagna di analisi dei dati. Il suo obiettivo è l'acquisizione integrale e fedele del dato grezzo dalla sorgente originaria, preservandone il contenuto informativo senza alcuna alterazione prematura.

Nel caso specifico del nostro progetto, il dataset *EEG.machinelearning_data_BRMH* è stato fornito in un singolo file in formato tabellare strutturato. Per avviare il processo di importazione, abbiamo utilizzato l'interfaccia nativa di Power BI Desktop, la quale, alla creazione di un nuovo progetto, presenta all'utente un pannello di connessione di dati che consente di selezionare tra molteplici sorgenti. Nella Figura 3.2 illustriamo la schermata iniziale di Power BI Desktop, nella quale sono visibili le opzioni di importazione disponibili.

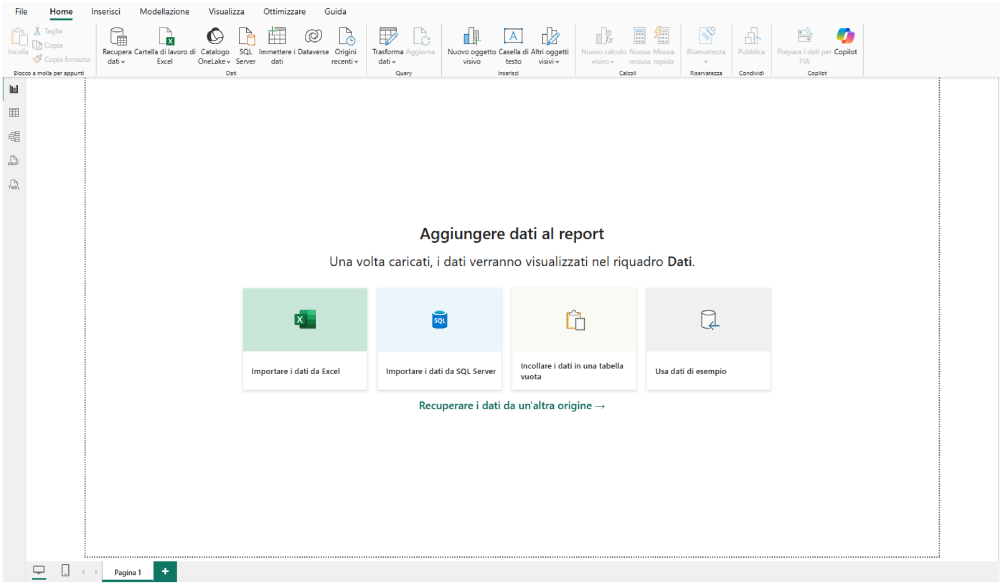
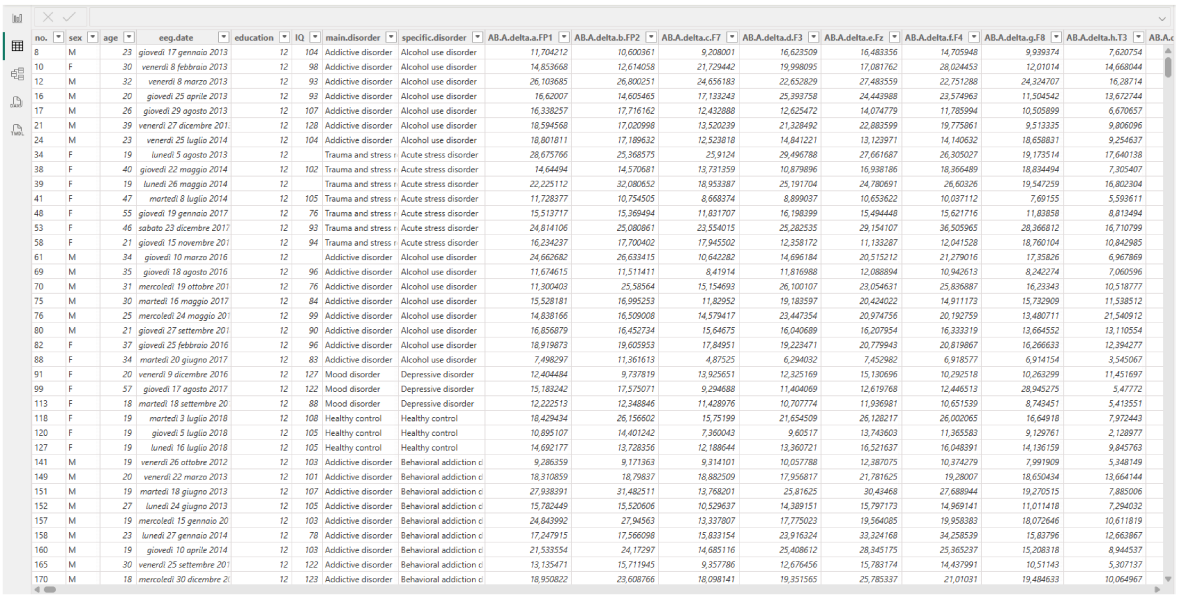


Figura 3.2: Schermata iniziale di Power BI Desktop con le opzioni di importazione di dati.

Una volta completato il caricamento del file sorgente, Power BI ha reso disponibile l'anteprima del dataset nella vista tabellare, permettendoci di effettuare una prima ispezione visiva del dato grezzo. Nella Figura 3.3 mostriamo un estratto di tale anteprima, nel quale risultano immediatamente visibili le colonne anagrafiche, diagnostiche e le prime variabili spettrali relative alla banda Delta.



no.	sex	age	eeg.date	education	IQ	main.disorder	specific.disorder	AB.A.delta.a.FP1	AB.A.delta.b.FP2	AB.A.delta.c.F7	AB.A.delta.d.F3	AB.A.delta.e.Fz	AB.A.delta.f.F4	AB.A.delta.g.F8	AB.A.delta.h.T3	AB.A.c
8	M	23	giovedì 17 gennaio 2013	12	104	Addictive disorder	Alcohol use disorder	11,704212	10,600361	9,208001	16,623509	16,483356	14,705948	9,939374	7,620754	
10	F	30	venerdì 8 febbraio 2013	12	98	Addictive disorder	Alcohol use disorder	14,853668	12,614058	21,729442	19,998095	17,081762	28,024453	12,01014	14,668044	
12	M	32	venerdì 8 marzo 2013	12	93	Addictive disorder	Alcohol use disorder	26,103685	26,800251	24,656183	22,652829	27,483559	22,751288	24,324707	16,28714	
16	M	20	giovedì 25 aprile 2013	12	93	Addictive disorder	Alcohol use disorder	16,62007	14,605465	17,133243	25,393758	24,449888	23,574963	11,504542	13,672744	
17	M	26	giovedì 29 agosto 2013	12	107	Addictive disorder	Alcohol use disorder	16,338257	17,716162	12,432888	12,625472	14,074779	11,785994	10,505899	6,670657	
21	M	39	venerdì 27 dicembre 201	12	128	Addictive disorder	Alcohol use disorder	18,594568	17,020998	13,520239	21,328492	22,883599	19,775861	9,513335	9,806096	
24	M	23	venerdì 25 luglio 2014	12	104	Addictive disorder	Alcohol use disorder	18,801811	17,189632	12,523818	14,841221	13,123971	14,140632	18,658831	9,254637	
34	F	19	lunedì 5 agosto 2013	12		Trauma and stress r	Acute stress disorder	28,675766	25,368575	25,9124	29,496788	27,661687	26,305027	19,173514	17,640138	
38	F	40	giovedì 22 maggio 2014	12	102	Trauma and stress r	Acute stress disorder	14,64494	14,570681	13,731359	10,879896	16,938186	18,366489	18,834494	7,305407	
39	F	19	lunedì 26 maggio 2014	12		Trauma and stress r	Acute stress disorder	22,225112	32,080652	18,953387	25,191704	24,780691	26,60326	19,547259	16,802304	
41	F	47	martedì 8 luglio 2014	12	105	Trauma and stress r	Acute stress disorder	11,728377	10,754505	8,668374	8,899037	10,653622	10,037112	7,69155	5,593611	
48	F	55	giovedì 19 gennaio 2017	12	76	Trauma and stress r	Acute stress disorder	15,513717	15,369494	11,831707	16,198399	15,494448	15,621716	11,83858	8,813494	
53	F	46	sabato 23 dicembre 2017	12	93	Trauma and stress r	Acute stress disorder	24,814106	25,080861	23,554015	25,282535	29,154107	36,505965	28,366812	16,710799	
58	F	21	giovedì 15 novembre 201	12	94	Trauma and stress r	Acute stress disorder	16,234237	17,700402	17,945502	12,358172	11,133287	12,041528	18,760104	10,842985	
61	M	34	giovedì 10 marzo 2016	12		Addictive disorder	Alcohol use disorder	24,662682	26,633415	10,642282	14,696184	20,515212	21,279016	17,35826	6,967869	
69	M	35	giovedì 18 agosto 2016	12	96	Addictive disorder	Alcohol use disorder	11,674615	11,511411	8,41914	11,816988	12,088894	10,942613	8,242274	7,060596	
70	M	31	mercoledì 19 ottobre 201	12	76	Addictive disorder	Alcohol use disorder	11,300403	25,58564	15,154693	26,100107	23,054631	25,836887	16,23343	10,518777	
75	M	30	martedì 16 maggio 2017	12	84	Addictive disorder	Alcohol use disorder	15,528181	16,995253	11,82952	19,183597	20,424022	14,911173	15,732909	11,538512	
76	M	25	mercoledì 24 maggio 201	12	99	Addictive disorder	Alcohol use disorder	14,838166	16,509008	14,579417	23,447354	20,974756	20,192759	13,480711	21,540912	
80	M	21	giovedì 27 settembre 201	12	90	Addictive disorder	Alcohol use disorder	16,856879	16,452734	15,64675	16,040689	16,207954	16,333219	13,664552	13,110554	
82	F	37	giovedì 25 febbraio 2016	12	96	Addictive disorder	Alcohol use disorder	18,919873	19,605953	17,84951	19,223471	20,779943	20,819867	16,266633	12,394277	
88	F	34	martedì 20 giugno 2017	12	83	Addictive disorder	Alcohol use disorder	7,498297	11,361613	4,87525	6,294032	7,452982	6,918577	6,914154	3,545067	
91	F	20	venerdì 9 dicembre 2016	12	127	Mood disorder	Depressive disorder	12,404484	9,737819	13,925651	12,325169	15,130696	10,292518	10,263299	11,451697	
99	F	57	giovedì 17 agosto 2017	12	122	Mood disorder	Depressive disorder	15,183242	17,575071	9,294688	11,404069	12,619768	12,446513	28,945275	5,47772	
113	F	18	martedì 18 settembre 20	12	88	Mood disorder	Depressive disorder	12,222513	12,348846	11,428976	10,707774	11,936981	10,651539	8,743451	5,413551	
118	F	19	martedì 3 luglio 2018	12	108	Healthy control	Healthy control	18,429434	26,156602	15,75199	21,654509	26,128217	26,002065	16,64918	7,972443	
120	F	19	giovedì 5 luglio 2018	12	105	Healthy control	Healthy control	10,895107	14,401242	7,360043	9,60517	13,743603	11,365583	9,129761	2,128977	
127	F	19	lunedì 16 luglio 2018	12	105	Healthy control	Healthy control	14,692177	13,728356	12,188644	13,360721	16,521637	16,048391	14,136159	9,845763	
141	M	19	venerdì 26 ottobre 2012	12	103	Addictive disorder	Behavioral addiction d	9,286359	9,171363	9,314101	10,057788	12,387075	10,374279	7,991909	5,348149	
149	M	20	venerdì 22 marzo 2013	12	101	Addictive disorder	Behavioral addiction d	18,310859	18,79837	18,882509	17,956817	21,781625	19,28007	18,650434	13,664144	
151	M	19	martedì 18 giugno 2013	12	107	Addictive disorder	Behavioral addiction d	27,938391	31,482511	13,768201	25,81625	30,43468	27,688944	19,270515	7,885006	
152	M	27	lunedì 24 giugno 2013	12	105	Addictive disorder	Behavioral addiction d	15,782449	15,520606	10,529637	14,389151	15,797173	14,969141	11,011418	7,294032	
157	M	19	mercoledì 15 gennaio 20	12	103	Addictive disorder	Behavioral addiction d	24,843992	27,94563	13,337807	17,775023	19,564085	19,958383	18,072646	10,611819	
158	M	23	lunedì 27 gennaio 2014	12	78	Addictive disorder	Behavioral addiction d	17,247915	17,566098	15,833154	23,916324	33,324168	34,258539	15,83796	12,663867	
160	M	19	giovedì 10 aprile 2014	12	103	Addictive disorder	Behavioral addiction d	21,533554	24,17297	14,685116	25,408612	28,345175	25,365237	15,208318	8,944537	
165	M	30	venerdì 25 settembre 201	12	122	Addictive disorder	Behavioral addiction d	13,135471	15,711945	9,357786	12,676456	15,783174	14,437991	10,51143	5,307137	
170	M	18	mercoledì 30 dicembre 2(	12	123	Addictive disorder	Behavioral addiction d	18,950822	23,608766	18,098141	19,351565	25,785337	21,01031	19,484633	10,064967	

Figura 3.3: Anteprima del dataset grezzo importato in Power BI, con le colonne anagrafiche, cliniche e le prime variabili spettrali della banda Delta.

Dall'analisi preliminare della qualità del dato grezzo sono emerse diverse criticità strutturali e di formato che hanno reso necessario un intervento mirato nella successiva fase di trasformazione. Innanzitutto, l'eterogeneità intrinseca dei sistemi di estrazione medica ha generato problematiche legate all'encoding dei caratteri testuali e all'inconsistenza dei separatori decimali (utilizzo misto di virgole e punti), un problema classico nell'importazione di CSV internazionali che Power BI richiede di risolvere definendo esplicitamente il *locale* di origine. In secondo luogo, la colonna identificativa del paziente (`no.`) veniva interpretata automaticamente dal motore di importazione come valore numerico intero, esponendola al rischio di aggregazioni matematiche accidentali (somme, medie) prive di significato semantico su un codice nominale. Inoltre, le 95 colonne contenenti le misurazioni di potenza spettrale presentavano una nomenclatura estremamente tecnica e compressa, del tipo `AB.A.delta.a.FP1`, nella quale il prefisso codificava la banda di frequenza e il suffisso identificava l'elettrodo secondo la convenzione del sistema internazionale 10-20. Infine, le variabili continue, come l'età e gli anni di scolarizzazione, pur risultando corrette nei valori grezzi, necessitavano di una discretizzazione categorica per poter alimentare adeguatamente i filtri interattivi delle dashboard.

3.5 Fase di trasformazione e pulizia

La fase di trasformazione rappresenta il nucleo operativo più corposo e strategicamente rilevante dell'intero processo. Al suo interno, il dato grezzo viene sottoposto a una catena di operazioni volte a sanificarlo, arricchirlo e renderlo semanticamente pronto per le analisi successive. Nel nostro progetto, le trasformazioni si sono articolate su due livelli distinti ma complementari: le operazioni eseguite all'interno dell'editor di Power Query prima del

caricamento nel modello, e le colonne calcolate e le misure definite successivamente tramite il linguaggio DAX.

3.5.1 Trasformazioni in Power Query

Il primo intervento ha riguardato la neutralizzazione dell'identificativo univoco del paziente. La colonna `no.`, originariamente interpretata come valore numerico intero, è stata forzata al formato testo. Questa operazione, apparentemente banale, costituisce, in realtà, una fondamentale *best practice* nella modellazione dei dati: essa inibisce qualsiasi aggregazione matematica accidentale su codici nominali, preservando la natura categorica dell'identificatore e impedendo che Power BI ne calcoli somme o medie prive di qualsiasi significato clinico.

Il secondo intervento ha riguardato l'ingegnerizzazione di una nuova variabile categorica a partire dal dato continuo dell'età anagrafica. Mediante logica condizionale implementata direttamente nell'editor di Power Query, la variabile `age` è stata trasformata nella colonna derivata *Fascia d'Età*, definendo quattro intervalli discreti con finalità di segmentazione epidemiologica: Under 25, 26-40, 41-60 e Over 60. Questa discretizzazione ha consentito di alimentare i filtri interattivi delle dashboard con categorie immediatamente comprensibili, facilitando l'analisi comparativa tra gruppi demografici omogenei.

Analogamente, la variabile continua `education`, che esprime gli anni di istruzione formale completati dal paziente, è stata trasformata nella colonna categorica *Fascia di Istruzione*. Come illustrato nella Figura 3.4, tale colonna suddivide il campione in due classi sulla base di una soglia clinicamente significativa, fissata a tredici anni di scolarizzazione. La formula DAX utilizzata per questa trasformazione è riportata nel Listato 3.1.

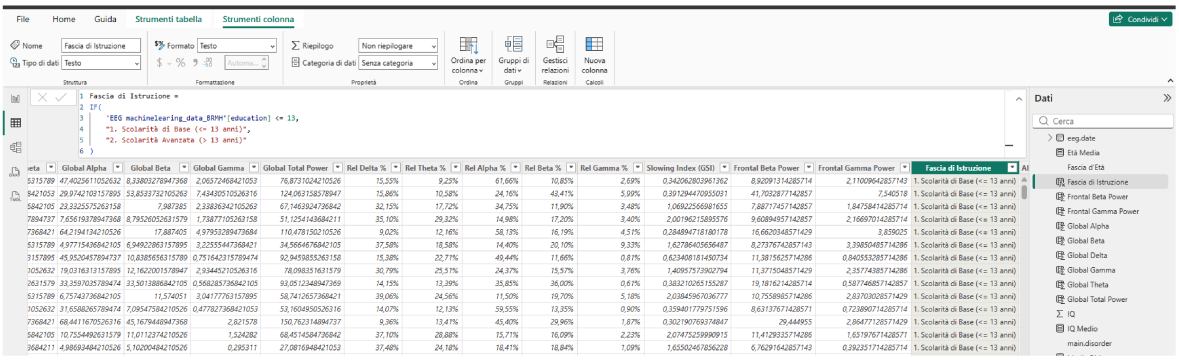


Figura 3.4: Schermata di Power BI durante la creazione della colonna calcolata *Fascia di Istruzione*, con la formula DAX visibile nella barra delle formule e il pannello Dati sulla destra.

```
Fascia di Istruzione =
IF(
    'EEG machinelearing_data_BRMH'[education] <= 13,
    "1. Scolarita' di Base (<= 13 anni)",
    "2. Scolarita' Avanzata (> 13 anni)"
)
```

Listing 3.1: Creazione della colonna calcolata *Fascia di Istruzione* mediante logica condizionale DAX.

3.5.2 Feature Engineering in DAX: razionale tecnico

Mentre Power Query è stato utilizzato per modellare la struttura e pulire il dato grezzo, il linguaggio DAX (Data Analysis Expressions) è stato impiegato per la successiva fase

di *feature engineering*. In questa fase abbiamo tradotto le misurazioni fisiche (i microvolt registrati dai singoli elettrodi) in veri e propri biomarcatori neurofisiologici. Dal punto di vista ingegneristico, è fondamentale distinguere tra le due tipologie di oggetti creati in DAX: le *Misure* e le *Colonne Calcolate*.

Le Misure operano sfruttando il cosiddetto *Filter Context* (contesto di filtro): esse non aggiungono dati fisici al modello, ma calcolano aggregazioni al volo in base ai filtri che l'utente applica nelle dashboard. Le Colonne Calcolate, invece, operano nel *Row Context* (contesto di riga): esse vengono valutate riga per riga al momento del caricamento e si materializzano fisicamente nel dataset, diventando nuove feature a tutti gli effetti. Poiché il nostro obiettivo era esplorare la varianza spaziale dell'EEG *paziente per paziente*, abbiamo implementato quasi tutte le metriche neurofisiologiche come Colonne Calcolate, riservando le Misure ai soli KPI anagrafici aggregati. Le motivazioni cliniche e teoriche che giustificano le formule di seguito presentate verranno esplorate in dettaglio nel Capitolo 4; in questa sede ci concentreremo sulla loro architettura computazionale.

3.5.3 Metriche di base (Misure)

Le prime tre metriche implementate sono Misure aggregate, progettate per restituire valori di coorte adattivi:

```
Pazienti Totali = COUNTROWS('EEG machinelearing_data_BRMH')
```

Listing 3.2: Misura DAX per il conteggio dinamico dei pazienti nel contesto filtrato.

```
Eta' Media = AVERAGE('EEG machinelearing_data_BRMH'[age])
```

Listing 3.3: Misura DAX per il calcolo dell'età media della coorte.

```
IQ Medio = AVERAGE('EEG machinelearing_data_BRMH'[IQ])
```

Listing 3.4: Misura DAX per il calcolo del quoziente intellettivo medio.

La funzione `AVERAGE` in DAX, a differenza di banali sommatorie, gestisce intrinsecamente i valori *null* o *blank* (come i pazienti per i quali il test del QI non è stato somministrato), escludendoli sia dal numeratore che dal denominatore e garantendo la purezza statistica del risultato.

3.5.4 Potenza globale e normalizzazione (Colonne Calcolate)

Per ridurre la dimensionalità del dataset (19 elettrodi per 5 bande = 95 variabili continue), abbiamo aggregato spazialmente l'attività corticale calcolando la potenza globale assoluta per ogni banda. Essendo una feature del singolo paziente, l'implementazione è avvenuta tramite Colonna Calcolata.

```
Global Alpha =
(
    'EEG machinelearing_data_BRMH'[AB.C.alpha.a.FP1] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.b.FP2] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.c.F7] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.d.F3] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.e.Fz] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.f.F4] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.g.F8] +
```

```
    'EEG machinelearing_data_BRMH'[AB.C.alpha.h.T3] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.i.C3] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.j.Cz] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.k.C4] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.l.T4] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.m.T5] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.n.P3] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.o.Pz] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.p.P4] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.q.T6] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.r.O1] +
    'EEG machinelearing_data_BRMH'[AB.C.alpha.s.O2]
) / 19
```

Listing 3.5: Colonna calcolata per la potenza globale della banda Alpha (media dei 19 elettrodi).

Lo stesso calcolo è stato replicato per *Global Delta*, *Global Theta*, *Global Beta* e *Global Gamma*. Sommandole, si ottiene la *Global Total Power*.

Poiché lo spettro EEG è fisiologicamente dominato dalle basse frequenze (legge $1/f$), la visualizzazione diretta dei valori assoluti produrrebbe grafici distorti e illeggibili. Per superare questa limitazione matematica, abbiamo ingegnerizzato la potenza relativa percentuale (la frazione di energia apportata da una singola banda rispetto al totale), utilizzando la funzione `DIVIDE` per scongiurare errori computazionali in caso di potenze totali nulle:

```
Rel Gamma % = DIVIDE(
    'EEG machinelearing_data_BRMH'[Global Gamma],
    'EEG machinelearing_data_BRMH'[Global Total Power]
)
```

Listing 3.6: Colonna calcolata per la potenza relativa percentuale della banda Gamma.

3.5.5 Biomarcatori sintetici e topologici

L'ultimo strato di *feature engineering* ha visto la traduzione in DAX di complessi indici clinici. Il *Global Slowing Index* (GSI) mette in rapporto le potenze relative delle onde lente e delle onde veloci, rivelando stati di rallentamento patologico:

```
Slowing Index (GSI) =
DIVIDE(
    ('EEG machinelearing_data_BRMH'[Rel Delta %] +
     'EEG machinelearing_data_BRMH'[Rel Theta %]),
    ('EEG machinelearing_data_BRMH'[Rel Alpha %] +
     'EEG machinelearing_data_BRMH'[Rel Beta %])
)
```

Listing 3.7: Colonna calcolata per il *Global Slowing Index*.

Il *Theta/Beta Ratio* (TBR), utilizzato clinicamente come indice di stress esecutivo, isola, invece, l'attività sull'elettrodo mediano frontale (Fz). Il terzo parametro della funzione `DIVIDE` (`BLANK()`) garantisce la robustezza visiva del modello restituendo un valore vuoto qualora il denominatore sia nullo.

```
Theta Beta Ratio (TBR) =
DIVIDE(
    'EEG machinelearing_data_BRMH'[AB.B.theta.e.Fz],
    'EEG machinelearing_data_BRMH'[AB.D.beta.e.Fz],
    BLANK()
)
```

Listing 3.8: Colonna calcolata per il *Theta/Beta Ratio* frontale.

Un'altra metrica fondamentale ingegnerizzata per le analisi topologiche è l'Indice di Asimmetria Alpha Frontale (FAA, *Frontal Alpha Asymmetry*). Questo biomarcatore, ampiamente documentato per lo studio dell'avvicinamento/evitamento emotivo, viene calcolato matematicamente come la differenza logaritmica naturale tra l'attività Alpha dell'emisfero destro (elettrodo F4) e quella dell'emisfero sinistro (elettrodo F3). La traduzione in linguaggio DAX impiega la funzione matematica `LN`:

```
Frontal Alpha Asymmetry (FAA) =
LN('EEG machinelearing_data_BRMH'[AB.C.alpha.f.F4]) - LN('EEG
    machinelearing_data_BRMH'[AB.C.alpha.d.F3])
```

Listing 3.9: Colonna calcolata per l'Asimmetria Alpha Frontale (FAA).

Infine, sono state estratte feature puramente topologiche come la *Frontal Beta Power* e la *Frontal Gamma Power*, mediando esclusivamente il segnale dei 7 elettrodi anteriori. Questo calcolo localizzato è il preludio alle analisi sull'efficienza neurale (che tratteremo nel Capitolo 4), in quanto isola computazionalmente l'attività delle aree cerebrali preposte alle funzioni esecutive superiori.

```
Frontal Gamma Power =
(
    'EEG machinelearing_data_BRMH'[AB.F.gamma.a.FP1] +
    'EEG machinelearing_data_BRMH'[AB.F.gamma.b.FP2] +
    'EEG machinelearing_data_BRMH'[AB.F.gamma.c.F7] +
    'EEG machinelearing_data_BRMH'[AB.F.gamma.d.F3] +
    'EEG machinelearing_data_BRMH'[AB.F.gamma.e.Fz] +
    'EEG machinelearing_data_BRMH'[AB.F.gamma.f.F4] +
    'EEG machinelearing_data_BRMH'[AB.F.gamma.g.F8]
) / 7
```

Listing 3.10: Colonna calcolata per la potenza Gamma frontale (7 elettrodi).

3.6 Fase di caricamento e modellazione

Una volta completata la fase di trasformazione e arricchimento dei dati tramite Power Query e le colonne calcolate DAX, il dataset processato è stato caricato definitivamente all'interno del motore analitico di Power BI. Questo passaggio, apparentemente automatico, riveste, in realtà, un'importanza architetturale significativa, poiché determina il modo in cui i dati vengono fisicamente memorizzati, indicizzati e resi disponibili per l'interrogazione in tempo reale da parte dei visual interattivi.

Il motore di memorizzazione sottostante a Power BI è VertiPaq, un database colonnare ottimizzato per operare interamente in memoria RAM. A differenza dei tradizionali motori relazionali basati su righe, VertiPaq memorizza i dati per colonna, applicando algoritmi di compressione avanzati che riducono drasticamente l'occupazione di memoria e consentono aggregazioni estremamente rapide anche su dataset di notevoli dimensioni. Nel nostro caso, il dataset finale, comprensivo delle colonne originali e di tutte le feature ingegnerizzate nella fase precedente, è stato caricato e compresso nel modello in memoria, pronto per alimentare le sei dashboard interattive descritte nei capitoli successivi.

Per quanto riguarda la modellazione relazionale, la struttura del nostro progetto si basa su un unico dataset monolitico che funge simultaneamente da tabella dei fatti e da contenitore delle dimensioni anagrafiche e cliniche. Questa scelta architetturale, pur discostandosi dal classico schema a stella descritto nel Capitolo 2, è risultata la più coerente con la natura del dataset clinico a disposizione: trattandosi di una singola tabella piatta già completa di tutte le informazioni necessarie (variabili anagrafiche, diagnostiche e spettrali per ciascun paziente), la creazione di tabelle delle dimensioni separate avrebbe introdotto complessità relazionale superflua senza apportare benefici analitici concreti. Le colonne derivate *Fascia d'Età* e *Fascia di Istruzione* fungono di fatto da dimensioni di segmentazione integrate direttamente nella tabella, alimentando i filtri interattivi senza necessità di tabelle di lookup esterne.

Infine, tutte le trasformazioni applicate in Power Query sono state salvate automaticamente sotto forma di passaggi sequenziali nello script del motore di trasformazione. Tale approccio garantisce la piena tracciabilità delle modifiche e consente di aggiornare automaticamente l'intero modello qualora il file sorgente venga modificato o arricchito con nuovi record, senza dover ripetere manualmente alcuna operazione.

CAPITOLO 4

Progettazione e realizzazione delle dashboard di base

L'obiettivo di questo capitolo è presentare in dettaglio la prima serie di dashboard sviluppate in Power BI. A differenza delle analisi puramente teoriche, in questa fase operiamo un'esplorazione pratica e diagnostica della coorte clinica, ponendo un forte accento sui biomarcatori neurofisiologici di base. Prima di illustrare le singole viste, il capitolo inquadra i principi teorici della Data Visualization in ambito biomedico, descrivendo l'approccio visivo adottato. Successivamente, per ciascuna dashboard esploriamo i meccanismi neurobiologici sottostanti, l'architettura degli oggetti visivi progettati e i risultati emersi dal campione studiato, validando le ipotesi di partenza.

4.1 Data Visualization e Business Intelligence nell'analisi clinica

La capacità umana di elaborare flussi informativi complessi è notevolmente amplificata dal supporto visivo, un paradigma che, nell'attuale scenario dei Big Data, diventa essenziale per la rapida comprensione dei fenomeni. Oggi, infatti, la ricerca clinica e biomedica è sommersa da enormi quantità di dati provenienti da fonti eterogenee, come referti clinici, test psicometrici e tracciati elettroencefalografici (EEG). La sfida principale consiste nell'esaminare queste informazioni per identificare rapidamente quelle realmente utili. La visualizzazione dei dati, o *Data Visualization*, risponde a questa esigenza, rendendo l'analisi molto più semplice e immediata. Con un solo sguardo a un grafico o a un diagramma è possibile individuare schemi ricorrenti, *outlier* patologici, tendenze e correlazioni.

4.1.1 Principi cognitivi della visualizzazione dei dati

Dal punto di vista cognitivo, la visualizzazione è estremamente efficace perché il 90% delle informazioni trasmesse al cervello è di natura visiva, e il nostro sistema nervoso elabora questi stimoli molto più velocemente rispetto al testo scritto. Ciò spiega perché la maggior parte dei ricercatori e dei medici reagisce in modo molto più reattivo a rappresentazioni grafiche rispetto a complesse tabelle contenenti centinaia di righe e decine di variabili (come nel caso dei 19 elettrodi per le 5 bande di frequenza del nostro dataset).

Pertanto, la *Data Visualization* non è soltanto uno strumento estetico di rappresentazione, ma un vero e proprio mezzo analitico per facilitare la comprensione profonda di dati biomedici complessi. Un grafico ben progettato non si limita a trasmettere un valore, ma ne amplifica l'impatto visivo, trasformando freddi numeri in *insight* clinici facilmente fruibili. Nella Figura

4.1 viene mostrata una panoramica concettuale di diversi tipi di *Data Visualization* applicati all'analisi avanzata dei dati.

Figura 4.1: Panoramica di diversi tipi di Data Visualization.

4.1.2 Power BI come strumento di visualizzazione clinica

Power BI è uno degli strumenti più diffusi e potenti per la visualizzazione e l'analisi dei dati. Esso consente di trasformare grandi volumi di dati, provenienti da fonti eterogenee, in *report* interattivi e *dashboard* dinamiche. Nel contesto del nostro lavoro di tesi neurofisiologica, l'adozione di Power BI rappresenta un salto di qualità metodologico: permette di superare i classici grafici statici stampati su carta o PDF, offrendo al lettore o al ricercatore la possibilità di "navigare" letteralmente all'interno delle onde cerebrali dei pazienti.

La caratteristica più rivoluzionaria di questa piattaforma è l'interattività nativa (*cross-filtering*): selezionando un singolo elemento in un grafico (ad esempio, la categoria "Schizofrenia" in un istogramma), tutti gli altri oggetti visivi della pagina (come grafici a dispersione o mappe spaziali) si riallineano e si ricalcolano istantaneamente per mostrare unicamente i dati di quel sottoinsieme. Questa capacità di operare un *drill-down* in tempo reale permette di isolare e confrontare fenotipi patologici che rimarrebbero altrimenti invisibili.

4.2 Gli oggetti visivi scelti per l'analisi neurofisiologica

L'efficacia di una visualizzazione dipende intrinsecamente dalla corretta selezione dell'oggetto visivo (*visual*) in relazione alla tipologia di dato che si intende rappresentare. Per tradurre i biomarcatori EEG in forme comprensibili, abbiamo adottato una specifica palette di strumenti grafici:

- *Mappa ad Albero (Treemap)*: utilizzata per mostrare gerarchie o scomposizioni di interi (es. il conteggio totale dei pazienti suddiviso per macro-patologia). Offre una percezione spaziale dei volumi molto più leggibile rispetto ai classici grafici a torta.
- *Grafico a Dispersione (Scatter Plot) e a Bolle*: lo strumento principe per individuare le correlazioni e gli *outlier*. Mappando le variabili continue sui due assi cartesiani (es. età e QI) e aggiungendo una terza dimensione tramite la dimensione della bolla, si svelano immediatamente i pattern o i cluster anomali tipici di specifiche disfunzioni cerebrali.
- *Grafico a Ragnatela (Radar Chart)*: estremamente efficace per confrontare *feature* multidimensionali che condividono la medesima scala. Nel nostro caso, è perfetto per visualizzare contemporaneamente le potenze relative di tutte e 5 le bande di frequenza (Delta, Theta, Alpha, Beta, Gamma) per una determinata patologia, disegnando una "firma" geometrica univoca.
- *Box Plot e Violin Plot*: strumenti statistici avanzati essenziali per non limitarsi al semplice valore medio. Mostrando la dispersione, la deviazione standard, i percentili e la densità di probabilità, certificano in modo trasparente se un valore clinico è omogeneo nella coorte o se presenta un'altissima varianza inter-soggetto.

L'insieme di queste architetture grafiche, combinate alle misure e alle trasformazioni descritte nel Capitolo 3, costituisce l'infrastruttura su cui poggiano le quattro dashboard di analisi descrittiva e diagnostica che verranno esaminate nelle sezioni seguenti.

4.3 Dashboard 1: Profilazione Clinica e Demografica

La prima dashboard, denominata "Profilazione Clinica e Demografica", rappresenta il punto di partenza metodologico del nostro studio. Nella Figura 4.2 illustriamo la schermata completa, progettata per offrire una visione macroscopica del panorama clinico e anagrafico della coorte in esame.

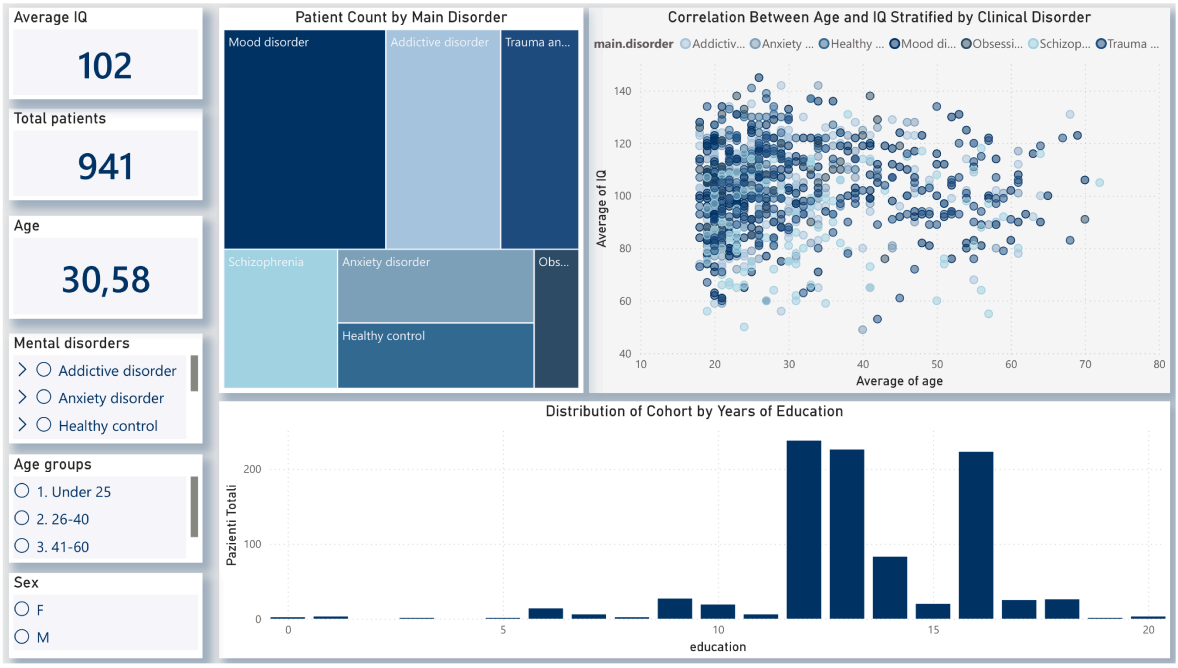


Figura 4.2: Schermata della Dashboard 1 relativa alla profilazione clinica e demografica della coorte.

4.3.1 Obiettivo e contesto analitico

L'obiettivo primario di questo primo pannello è eseguire un'attenta *Exploratory Data Analysis* (EDA). Prima di applicare algoritmi complessi sui segnali neuroelettrici, riteniamo fondamentale ottenere una comprensione ad alto livello della struttura epidemiologica della popolazione studiata. Questo approccio permette di validare la *Data Quality*, individuare eventuali *bias* anagrafici e valutare il bilanciamento delle classi diagnostiche.

In fase di ingegnerizzazione del dato tramite *Power Query*, abbiamo applicato alcune trasformazioni essenziali. L'identificativo univoco del paziente (la colonna `no.`) è stato forzato al formato testuale, applicando una *best practice* volta a inibire aggregazioni matematiche accidentali su codici nominali. Inoltre, per facilitare l'interazione visiva, la variabile continua relativa all'età è stata trasformata in una *feature* categorica, definendo la colonna `Fascia d'Età` con quattro raggruppamenti specifici: Under 25, 26-40, 41-60 e Over 60.

4.3.2 Descrizione degli oggetti visivi

L'architettura visiva è strutturata per minimizzare il carico cognitivo, presentando i dati chiave in modo intuitivo. Abbiamo inserito un pannello di filtri globale (*Slicer*) basato sulla variabile categorica `main.disorder`, che permette all'utente di isolare istantaneamente specifiche patologie psichiatriche o neurologiche. A supporto, tre schede KPI (*Key Performance Indicator*) restituiscono i valori anagrafici e cognitivi medi.

Per mostrare il volume dei pazienti suddiviso per macro-patologia, abbiamo optato per una Mappa ad Albero (*Treemap*). Questo oggetto visivo palesa immediatamente le proporzioni del campione, evidenziando le aree diagnostiche con maggiore densità rispetto ai classici, e spesso meno leggibili, diagrammi circolari.

Il nucleo analitico della dashboard è costituito dal Grafico a Dispersione (*Scatter Plot*). Questo grafico mappa la distribuzione bivariata dell'età (Asse X) rispetto al Quoziente Intellettivo (Asse Y). Parametrizzando il livello di dettaglio sul singolo paziente, ogni punto nel piano cartesiano rappresenta un'unità statistica, mentre il colore ne identifica la diagnosi.

Infine, un Istogramma a colonne descrive la distribuzione asimmetrica degli anni di scolarizzazione (`education`), una co-variata che terremo in considerazione nelle analisi successive riguardanti la riserva cognitiva.

4.3.3 Risultati e interpretazione

L'analisi di questa prima schermata certifica la presenza di un campione solido e statisticamente rilevante, costituito da 941 tracciati registrati, con un'età media che si attesta intorno ai 30,58 anni e un livello cognitivo globale pari a un QI medio di 102. Tali valori confermano un'ottima eterogeneità di base.

La Mappa ad Albero ha rivelato le patologie maggiormente rappresentate nel database, con una concentrazione significativa nei disturbi dell'umore, dell'ansia e dello spettro schizofrenico. Questo sbilanciamento volumetrico, noto come *class imbalance*, richiede particolare attenzione qualora i dati vengano impiegati per addestrare modelli predittivi, sebbene garantisca, al contempo, una ricchezza informativa per l'estrazione di *feature* psichiatriche.

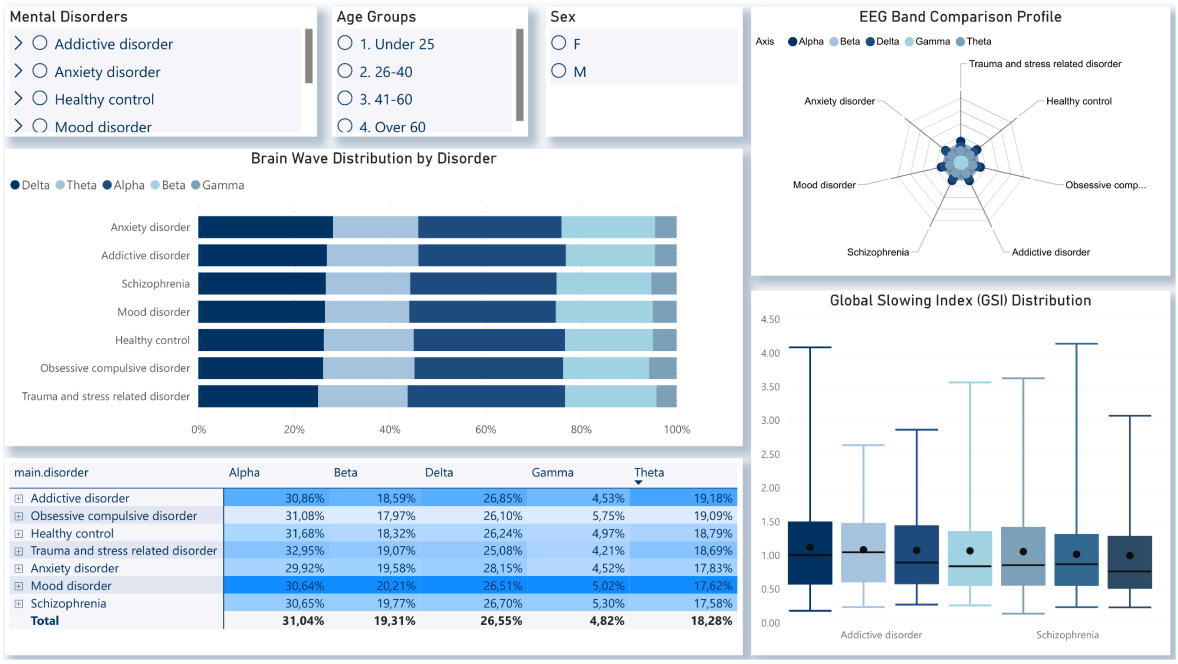
Il grafico a dispersione si è rivelato cruciale per l'individuazione di *outlier* clinici. Mappando in modo granulare l'interazione tra età e intelligenza, è possibile notare visivamente come alcune specifiche patologie tendano a concentrarsi in fasce di quoziente intellettivo nettamente inferiori, discostandosi dalla nuvola di punti principale dei soggetti di controllo.

L'istogramma sull'istruzione, infine, ha evidenziato picchi marcati intorno ai 12 e ai 16 anni di scolarizzazione. Aver tracciato questo parametro è essenziale: poiché l'istruzione contribuisce alla riserva cognitiva del cervello, conoscerne la distribuzione assicura che le

future alterazioni delle reti neurali non vengano erroneamente attribuite in via esclusiva alla patologia, ma siano contestualizzate rispetto al *background* del paziente.

4.4 Dashboard 2: Firme Spettrali e Global EEG Slowing

Questa sezione introduce l'analisi del segnale EEG nel dominio della frequenza. Nella Figura 4.3 riportiamo la Dashboard 2, focalizzata sull'estrazione delle firme spettrali e sulla quantificazione del livello di attivazione corticale globale.



main.disorder	Alpha	Beta	Delta	Gamma	Theta
Addictive disorder	30,86%	18,59%	26,85%	4,53%	19,18%
Obsessive compulsive disorder	31,08%	17,97%	26,10%	5,75%	19,09%
Healthy control	31,68%	18,32%	26,24%	4,97%	18,79%
Trauma and stress related disorder	32,95%	19,07%	25,08%	4,21%	18,69%
Anxiety disorder	29,92%	19,58%	28,15%	4,52%	17,83%
Mood disorder	30,64%	20,21%	26,51%	5,02%	17,62%
Schizophrenia	30,65%	19,77%	26,70%	5,30%	17,58%
Total	**31,04%**	**19,31%**	**26,55%**	**4,82%**	**18,28%**

Figura 4.3: Schermata della Dashboard 2 relativa all'estrazione delle firme spettrali e al Global EEG Slowing.

4.4.1 Fondamento teorico: il rallentamento globale dell'EEG

L'interpretazione diagnostica dei tracciati si basa su un principio neuroscientifico fondamentale e trasversale nella psichiatria basata sull'EEG: il *Global EEG Slowing* (Rallentamento Globale dell'EEG). In un cervello sano, in condizioni di veglia ad occhi chiusi, l'attività elettrica è tipicamente dominata dal ritmo Alpha (8-12 Hz) nelle regioni posteriori, segno di una corteccia "a riposo", ma metabolicamente attiva e pronta a reagire agli stimoli.

Tuttavia, in presenza di patologie psichiatriche e neurologiche gravi — come la schizofrenia cronica, la demenza, i traumi cranici (TBI) e le forme di depressione maggiore — si osserva un'alterazione strutturale e funzionale delle reti talamo-corticali. Questo deterioramento si traduce elettricamente in un aumento diffuso e massivo delle onde lente (bande Delta, 0.5-4 Hz, e Theta, 4-8 Hz) sull'intero scalpo, a totale discapito delle onde veloci (bande Alpha e Beta). Da un punto di vista clinico e cellulare, l'invasione delle frequenze Delta e Theta nello stato di veglia indica una ridotta efficienza metabolica, un abbassamento del flusso sanguigno cerebrale regionale e una generale ipoattivazione del cervello, in cui i neuroni faticano a mantenere lo stato di *arousal* (attivazione) necessario per le normali funzioni esecutive.

Per quantificare matematicamente e oggettivare questo fenomeno, la letteratura scientifica fa ampio ricorso al *Global Slowing Index* (GSI). Questo biomarcatore sintetico calcola il rapporto

scalare tra l'energia totale delle onde lente e l'energia delle onde veloci:

$$\text{GSI} = \frac{\text{Delta} + \text{Theta}}{\text{Alpha} + \text{Beta}} \tag{4.4.1}$$

Un valore basso di GSI denota un cervello reattivo e in stato di normale *arousal*, dove predominano i ritmi rapidi. Un valore GSI criticamente elevato funge invece da spartiacque diagnostico, evidenziando in modo inequivocabile un rallentamento patologico della conduzione neurale e uno stato di neurodegenerazione o desincronizzazione cronica.

4.4.2 Descrizione degli oggetti visivi

Al fine di isolare questo biomarcatore, l'infrastruttura dati è stata arricchita con nuove metriche calcolate in DAX. È stata, innanzitutto, calcolata la *Global Total Power*, ovvero l'energia complessiva media su tutti i 19 elettrodi, utilizzata successivamente per normalizzare la potenza di ogni singola banda (Potenza Relativa). Tale normalizzazione percentuale è uno step algoritmico vitale per impedire che il rumore fisiologico intrinseco dei segnali biologici (la cosiddetta legge $1/f$ o rumore rosa) appiattisca le visualizzazioni, garantendo un confronto puro tra le diverse componenti in scala $0 - 100\%$.

Gli oggetti visivi scelti per la schermata rispondono all'esigenza di confrontare i profili energetici delle diverse patologie. Un grafico a barre in pila al 100% rappresenta la scomposizione della potenza relativa nelle cinque bande di frequenza in funzione della macro-diagnosi. Un Grafico a Ragnatela (*Radar Chart*) traduce le potenze relative in una firma geometrica vettoriale, permettendo di sovrapporre e confrontare visivamente i profili energetici tracciando cinque linee di tensione (una per banda).

A supporto dell'analisi di dettaglio, una Matrice con formattazione condizionale a mappa di calore incrocia le diagnosi principali e specifiche con le potenze relative, accendendo cromaticamente i picchi di ipo o iper-attivazione. Infine, un *Box Plot* posizionato in basso a destra descrive la distribuzione statistica e la dispersione del GSI per ciascuna categoria clinica. È importante sottolineare che, per aggregare i dati nei grafici spettrali, è stata utilizzata la media aritmetica e non la somma vettoriale, il che ha permesso di collassare la varianza inter-soggetto e di restituire l'esatto profilo cerebrale medio tipico della classe analizzata.

4.4.3 Risultati e interpretazione

L'analisi dei descrittori globali restituisce un GSI medio della coorte pari a 1,42, un valore che funge da baricentro del dataset. L'efficacia della normalizzazione emerge chiaramente dal grafico a barre, dove patologie gravi, come la demenza o la schizofrenia, presentano un'espansione macroscopica delle fasce relative alle frequenze Delta e Theta, le quali arrivano a occupare una porzione prevalente (oltre il 60%) dell'intera energia cerebrale, comprimendo quasi totalmente le onde rapide deputate ai processi cognitivi superiori.

Tale evidenza trova riscontro immediato e intuitivo nel Grafico a Ragnatela: in corrispondenza del vertice relativo ai controlli sani si nota una forte estensione del vettore Alpha; al contrario, le figure geometriche corrispondenti ai disturbi severi subiscono un netto stiramento geometrico verso la periferia del grafico in direzione dei vertici delle onde lente, offrendo una prova visiva immediata della ridotta efficienza corticale.

L'analisi del *Box Plot* fornisce la validazione statistica decisiva, superando i limiti della media matematica semplice. Il box relativo ai soggetti sani si posiziona nella parte bassa dell'asse Y ed è molto compatto, denotando stabilità e omogeneità nella coorte di controllo. Le "scatole" relative a patologie come la schizofrenia, invece, non solo sono traslate verso l'alto sull'asse delle ordinate (a riprova dell'avvenuto rallentamento), ma risultano anche notevolmente allungate sull'asse verticale. Questa elevata varianza inter-soggetto testimonia

che la malattia si manifesta elettricamente con gradi di gravità estremamente eterogenei, generating numerosi *outlier* clinici. Individuare visivamente tali *outlier* è fondamentale, in quanto si tratta di pazienti con un indice di rallentamento critico che richiederanno strategie di regolarizzazione specifiche nella successiva fase di addestramento dei modelli predittivi.

4.5 Dashboard 3: Asse Cognitivo (Efficienza Neurale e Sincronizzazione Gamma)

Questa sezione illustra la Dashboard 3 (Figura 4.4), il cui focus si sposta dall'analisi globale dello scalpo all'indagine topologica mirata sui lobi frontali. L'obiettivo è esplorare le profonde correlazioni biologiche tra le metriche psicometriche di intelligenza e l'efficienza dei ritmi elettrici rapidi.

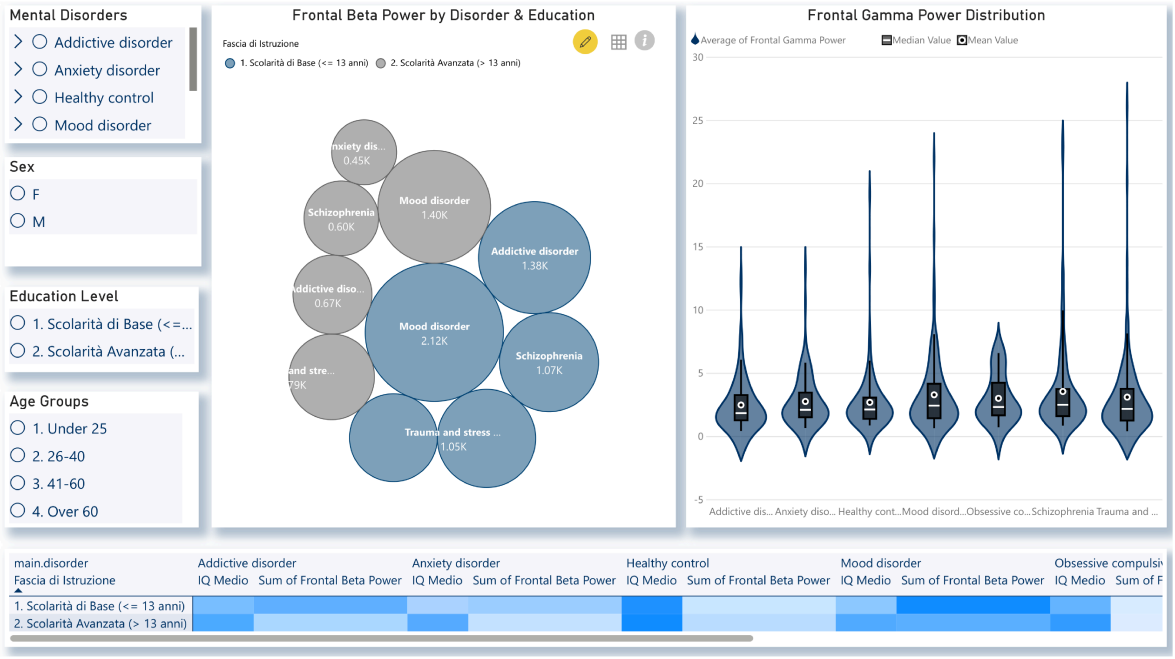


Figura 4.4: Schermata della Dashboard 3 relativa all'Asse Cognitivo.

4.5.1 Fondamento teorico: efficienza neurale e sincronizzazione Gamma

Questa dashboard testa empiricamente due pilastri della neuroscienza cognitiva moderna che smentiscono molte intuizioni comuni sul funzionamento cerebrale. Il primo è l'*Ipotesi dell'Efficienza Neurale*. Sebbene il senso comune suggerisca che un cervello più intelligente debba "lavorare di più", la letteratura scientifica ha dimostrato esattamente il contrario. L'ipotesi dell'efficienza neurale [Neubauer e Fink, 2009] postula che i soggetti dotati di un QI elevato consumino, a parità di *task*, meno energia metabolica globale rispetto a soggetti con QI inferiore. Tuttavia, questo risparmio energetico diffuso è bilanciato da una straordinaria capacità di generare una potenza di calcolo ad altissima frequenza (Banda Beta, 13-30 Hz) all'interno di reti corticali estremamente localizzate e specializzate, prime fra tutte i lobi frontali. In altre parole, l'intelligenza non risiede nella quantità bruta di attivazione corticale, ma nell'ottimizzazione del *routing* e nella focalizzazione delle risorse neurali laddove e quando il carico cognitivo lo richiede strettamente.

Il secondo principio fondamentale investigato è il *Binding Problem* (Problema del Legame) e la sincronizzazione della banda Gamma (> 30 Hz). Le onde Gamma rappresentano il ritmo

oscillatorio più veloce e metabolicamente costoso del cervello. A livello neurofisiologico, esse svolgono la funzione critica di "legare" informazioni sensoriali e concettuali elaborate e provenienti da aree anatomiche diverse in un unico e coerente percetto o pensiero cosciente. Sono, in sintesi, la colla della coscienza, fondamentali per l'apprendimento profondo e la memoria di lavoro. Nei soggetti affetti da gravi patologie psichiatriche, in particolare la schizofrenia, questo delicato meccanismo di sincronizzazione a lungo raggio risulta gravemente compromesso. Il deterioramento delle firme Gamma non comporta semplicemente una variazione quantitativa della frequenza, ma un collasso qualitativo del ritmo, portando alla frammentazione del pensiero, all'incapacità di processare coerentemente la realtà e, nei casi più gravi, all'insorgenza di deliri e allucinazioni.

4.5.2 Descrizione degli oggetti visivi

Per supportare e dimostrare visivamente queste teorie complesse, l'infrastruttura dati è stata estesa con l'ingegnerizzazione di nuove *feature* selettive. Tramite colonne calcolate, è stata isolata la componente topologica prefrontale, estraendo la media matematica delle bande Beta e Gamma esclusivamente dal cluster di elettrodi frontali (FP1, FP2, F3, F4, Fz, F7, F8). Inoltre, la variabile inerente agli anni di scolarizzazione è stata discretizzata in una colonna categorica, creando un *binning* culturale che separa la "Scolarità di Base" ($\leq$ 13 anni) dalla "Scolarità Avanzata" ($>$ 13 anni).

L'oggetto visivo principale è un Grafico a Bolle a Tre Dimensioni (*Bubble Chart*) che mappa l'interazione cartesiana continua tra il QI (Asse X) e la potenza Beta frontale (Asse Y), segmentando i pazienti per colore in base alla macro-diagnosi. Il grafico introduce una terza dimensione di analisi mappando l'ampiezza delle bolle sull'età del paziente. L'inserimento delle rette di regressione lineare formalizza visivamente la pendenza del gradiente di efficienza neurale.

Un secondo strumento analitico avanzato, il Grafico a Violino (*Violin Plot*), analizza la distribuzione statistica della potenza Gamma frontale. Integrando al suo interno la robustezza del *Box Plot* e all'esterno la fluidità della stima della densità kernel (KDE), il grafico svela la morfologia reale della distribuzione del segnale, superando i limiti dei semplici istogrammi. Infine, una Matrice di Controllo incrocia la fascia di istruzione con le macro-patologie, calcolando i valori medi di QI e Beta frontale supportati da una formattazione condizionale a mappa di calore.

4.5.3 Risultati e interpretazione

L'incrocio tra metriche psicometriche e tensori biologici evidenzia divergenze radicali tra le classi. L'analisi del grafico a bolle rivela un comportamento non uniforme delle linee di tendenza: nel gruppo dei controlli sani si osserva una pendenza definita e coerente, a testimonianza del fatto che i cervelli cognitivamente più performanti richiedano e riescano a erogare una sincronizzazione Beta frontale ottimizzata. Al contrario, nelle coorti affette da disturbi psicotici — con particolare evidenza per la schizofrenia — la retta di regressione subisce una severa flessione e appiattimento, palesando una rottura totale del gradiente lineare. Questo appiattimento indica una disconnessione strutturale tra il potenziale cognitivo (QI) e la sua effettiva espressione elettrica prefrontale.

Il Grafico a Violino svela, poi, l'anomala micro-struttura della banda Gamma. Mentre i soggetti sani mostrano una distribuzione standard, caratterizzata da una campana simmetrica e fortemente concentrata attorno alla mediana (indice di un ritmo Gamma stabile, integro e deputato al corretto *binding* informativo), il violino della schizofrenia e dei disturbi cognitivi gravi presenta una morfologia innaturalmente allungata e spesso bimodale. Questa distorsione geometrica visibile nella KDE dimostra matematicamente una sregolazione massiva delle

oscillazioni ad alta frequenza, fornendo una spiegazione biologica tangibile per i deficit della memoria di lavoro e la frammentazione del pensiero tipici di questi pazienti.

La Matrice di Controllo ha, infine, permesso di isolare e neutralizzare la presenza di eventuali *bias* legati alla riserva culturale. I risultati hanno evidenziato che la potenza Beta frontale è intrinsecamente legata alla patologia e al QI biologico, e non risente in modo esclusivo degli anni di studio accumulati. Questo dato garantisce un'assoluta robustezza analitica: se da un lato la riserva cognitiva indotta da una scolarità avanzata agisce innegabilmente come modulatore del punteggio nei test del QI, dall'altro le alterazioni organiche delle onde rapide frontali rimangono un'impronta neurobiologica incancellabile della patologia di fondo. L'estrazione di queste *feature* topologiche costituisce una base di altissimo valore predittivo per gli stadi di classificazione automatica.

4.6 Dashboard 4: Asimmetria Alpha Frontale (FAA)

A chiusura della suite esplorativa e diagnostica, la Dashboard 4 (Figura 4.5) opera un *drill-down* analitico sulla dimensione neurofisiologica della regolazione emotiva e dell'umore. L'architettura visiva abbandona l'analisi energetica globale e scompone un biomarcatore tensoriale estremamente specifico attraverso stratificazioni anagrafiche, biologiche e cliniche.

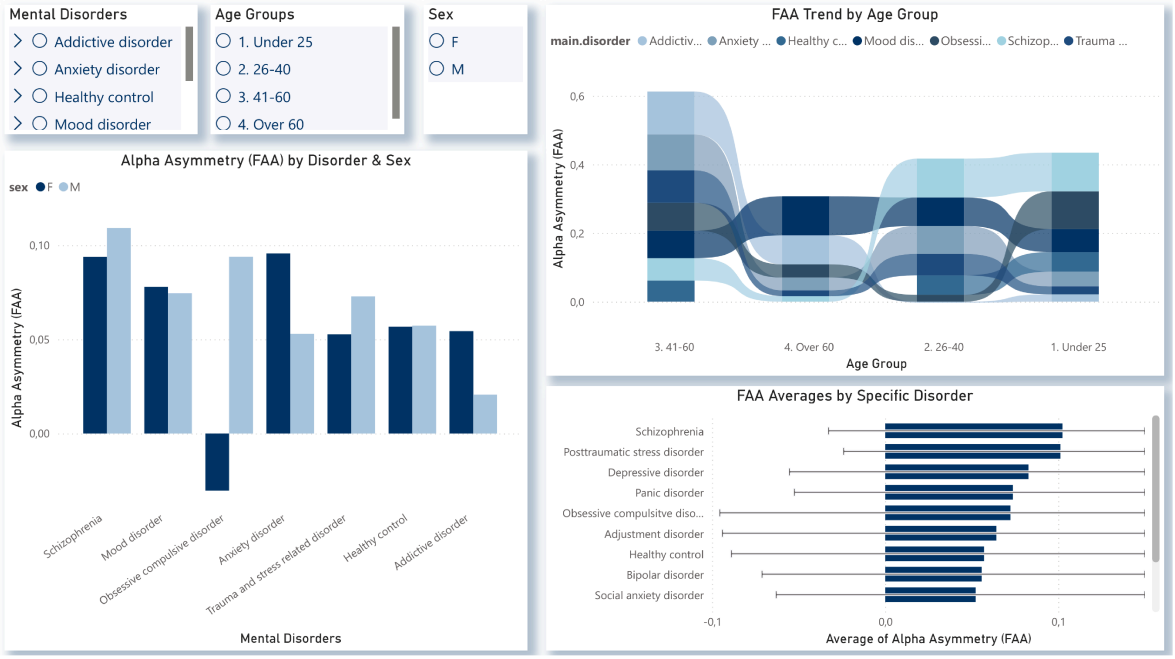


Figura 4.5: Schermata della Dashboard 4 relativa ai biomarcatori dell'umore e all'Asimmetria Alpha Frontale.

4.6.1 Fondamento teorico: la FAA e i modelli BIS/BAS

L'indagine oggettiva sulle basi organiche delle emozioni ha le sue radici nel lavoro pionieristico del neuroscienziato Richard J. Davidson. Nel corso degli ultimi decenni, la sua ricerca ha consolidato un solido modello di specializzazione emisferica frontale strettamente associata all'affettività [Davidson, 1992]. Secondo il Modello dell'Asimmetria Frontale di Davidson, l'attività corticale dell'emisfero frontale sinistro è intrinsecamente legata ai comportamenti di avvicinamento, alle emozioni positive e alla motivazione orientata a uno scopo. Viceversa,

un'attività corticale preponderante nell'emisfero frontale destro è biologicamente deputata a elaborare comportamenti di evitamento, ritiro sociale, allerta ed emozioni a valenza negativa.

Da un punto di vista psicologico e motivazionale, le scoperte empiriche del *dove* avvengano queste attivazioni (i lobi frontali) sono state mirabilmente teorizzate nel *perché* dallo psicologo Jeffrey Gray, creatore della Teoria della Sensibilità al Rinforzo [Gray, 1981]. Gray ha teorizzato l'esistenza di due complessi sistemi comportamentali che regolano la nostra risposta agli stimoli; ovvero:

- Il *Behavioral Activation System (BAS)*: il Sistema di Attivazione Comportamentale, ancorato elettricamente all'emisfero sinistro, incoraggia l'azione attraverso aspettative positive. Si concentra su ricompense, conquiste e genera sentimenti di speranza, euforia e motivazione.
- Il *Behavioral Inhibition System (BIS)*: il Sistema di Inibizione Comportamentale, ancorato all'emisfero destro, domina il processo decisionale quando i soggetti si trovano a dover evitare esperienze avverse o punizioni. Il BIS limita il comportamento e genera potenti emozioni inibitorie, come l'ansia cronica, la paura e la profonda tristezza.

Per misurare l'equilibrio tra questi due delicati sistemi, la bioinformatica contemporanea si affida all'Indice Globale di Asimmetria Alpha (FAA, *Frontal Alpha Asymmetry*). Poiché l'onda Alpha ha un carattere tipicamente inibitorio (più Alpha significa meno attivazione corticale sottostante), l'FAA è calcolato matematicamente come la differenza logaritmica dell'attività Alpha tra l'elettrodo frontale destro (F4) e il sinistro (F3):

$$\text{FAA} = \ln(F4) - \ln(F3) \tag{4.6.1}$$

Un valore FAA positivo indica un'attivazione maggiore dell'emisfero sinistro (minore inibizione Alpha a sinistra rispetto a destra), rivelando una dominanza del sistema BAS. Un valore FAA fortemente negativo indica, invece, un'asimmetria destrorsa, rivelando un'iperattivazione cronica del sistema BIS: questa condizione è considerata un *marker* primario per i disturbi depressivi severi e per l'ansia patologica.

4.6.2 Descrizione degli oggetti visivi

La Dashboard 4 elabora l'indice FAA attraverso un sistema di filtraggio avanzato in modalità *cross-filtering*, che consente di isolare specifiche sottopopolazioni a *runtime*. I filtri, basati su macro-diagnosi, fascia d'età e sesso biologico, ricalcolano istantaneamente l'aggregazione spaziale sui visual.

Il Grafico a Colonne Raggruppate, posizionato nella sezione superiore, incrocia la media dell'indice FAA con le macro-patologie, introducendo una suddivisione categorica per genere. Questo *split* cromatico (maschi vs femmine) è uno strumento diagnostico essenziale per isolare l'interazione critica tra sesso biologico e la manifestazione neuroelettrica del disturbo.

A sinistra, un Grafico a Barre Orizzontali mappa le diagnosi di secondo livello (diagnosi specifiche) rispetto al punteggio FAA. L'implementazione metodologicamente più rigorosa e qualificante di questa rappresentazione grafica è la sovrapposizione delle Barre di Errore (*Error Bars*). Tali barre permettono di visualizzare simultaneamente la tendenza centrale (la media aggregata) e gli intervalli di deviazione standard, fornendo l'effettiva dispersione clinica che una semplice media nasconderebbe matematicamente.

Infine, l'analisi della traiettoria anagrafica è demandata a un Grafico a Nastro (*Ribbon Chart*). Questo elaborato strumento elimina il grave errore metodologico dell'aggregazione a cascata, trattando la metrica come un flusso continuo nel tempo; esso ordina geometricamente dall'alto in basso le patologie all'interno di ogni fascia d'età sulla base della positività

dell'FAA, tracciando nastri fluidi di collegamento. Lo scavalcamento reciproco dei nastri (il cambio di rango) permette di identificare visivamente se l'asimmetria Alpha tenda a convergere verso la normalità con l'invecchiamento cerebrale, oppure se esistano finestre critiche di vulnerabilità biologica specifiche per ogni macro-disturbo.

4.6.3 Risultati e interpretazione

Il Grafico a Colonne Raggruppate ha immediatamente svelato un *bias* di genere di enorme impatto clinico all'interno del campione. Analizzando la categoria del Disturbo Ossessivo-Compulsivo (OCD), emerge una divergenza radicale e contro-intuitiva: mentre i pazienti di sesso maschile affetti da OCD mantengono, nel campione, un punteggio FAA fortemente positivo (dominanza sinistra e sistema BAS), il costrutto misurato per i pazienti di sesso femminile con la medesima diagnosi precipita in modo vertiginoso nel quadrante negativo (FAA < 0). Questa asimmetria destrorsa unicamente femminile palesa una grave iperattivazione del sistema BIS (ansia, evitamento degli stimoli e rimuginio intrusivo), confermando in modo empirico che la medesima etichetta diagnostica psichiatrica può mascherare e manifestarsi con architetture elettriche diametralmente opposte a seconda del sesso biologico del paziente.

Il Grafico a Barre Orizzontali ha affrontato il paradosso statistico dell'appiattimento aggregato. A una prima lettura superficiale della tendenza centrale, quasi tutte le barre dei valori medi risultavano posizionate nel quadrante positivo, suggerendo un'apparente "salute" emotiva o motivazionale a livello globale. Tuttavia, l'implementazione delle barre di errore ha rivelato l'effettiva micro-struttura del dato: patologie cardine dell'affettività, come il Disturbo Depressivo e il Disturbo Bipolare, presentano code di dispersione che "bucano" ampiamente la linea dello zero, estendendosi in profondità nei valori negativi (fino a FAA $\approx -0,10$). Questa evidenza visiva e matematica ha smascherato l'esistenza di ampi *cluster* di pazienti affetti da sintomatologia depressiva severa e inibizione destrorsa, i quali risultavano silenziosamente cancellati dalla media algebrica di gruppo che neutralizzava poli opposti.

Nel complesso, l'infrastruttura analitica di questa dashboard convalida indiscutibilmente l'utilizzo dell'elettroencefalografia come formidabile strumento di fenotipizzazione psichiatrica oggettiva. Aver dimostrato e documentato visivamente l'esistenza delle asimmetrie negative isolate e delle forti divergenze funzionali di genere eleva la solidità del dataset, preparandolo in via definitiva per essere processato con efficacia dagli algoritmi di Intelligenza Artificiale per la classificazione predittiva affrontati nel capitolo successivo.

CAPITOLO 5

Progettazione e realizzazione delle dashboard avanzate

Il presente capitolo conclude il percorso analitico intrapreso nel Capitolo 4, portando l'indagine diagnostica a un livello di complessità superiore attraverso la progettazione di due dashboard avanzate. La prima, dedicata al Rapporto Theta/Beta (Theta/Beta Ratio, TBR), approfondisce il dominio della vulnerabilità allo stress e del controllo esecutivo frontale, introducendo un biomarcatore storicamente validato per la valutazione dell'equilibrio attentivo e dell'attivazione corticale. La seconda dashboard segna un cambio di paradigma radicale, integrando algoritmi di Intelligenza Artificiale — apprendimento non supervisionato tramite Clustering K-Means, apprendimento supervisionato tramite Regressione Logistica e analisi gerarchica della varianza — in un unico pannello interattivo a filtraggio incrociato. Il capitolo si conclude con la discussione del ciclo analitico iterativo che connette i tre modelli, trasformando il database statico in un motore di scoperta di insight biomedici.

5.1 Analisi avanzata in Power BI: strumenti e approcci

L'evoluzione dell'ecosistema di *Business Intelligence* ha progressivamente eroso il confine storico tra la reportistica descrittiva, tradizionalmente affidata agli analisti di business, e la modellazione predittiva, storicamente dominio esclusivo dei *data scientist*. Questa convergenza tecnologica trova la sua massima espressione nelle funzionalità di analisi avanzata integrate nativamente all'interno di Power BI. In questa prima sezione, si delinea il framework concettuale e tecnologico che abilita il passaggio verso forme di analisi superiori, gettando le basi teoriche per le implementazioni pratiche che verranno discusse nel prosieguo del capitolo.

5.1.1 Dall'analisi diagnostica all'analisi predittiva

Facendo riferimento al paradigma del ciclo di vita della *Data Analytics* introdotto nel Capitolo 1, il lavoro svolto finora si è concentrato prevalentemente sui primi due gradini della piramide del valore analitico: l'analisi descrittiva e l'analisi diagnostica. Attraverso le dashboard presentate nel Capitolo 4, abbiamo risposto alle domande "*Cosa è successo?*" (fotografando la distribuzione demografica e clinica del campione) e "*Perché è successo?*" (indagando le asimmetrie frontali o il rallentamento globale dell'EEG in relazione alle singole patologie).

Tuttavia, per sbloccare il vero potenziale di un dataset biomedico complesso, è necessario approdare all'analisi predittiva (*Predictive Analytics*). L'obiettivo di questa fase non è più

soltanto la sintesi e la visualizzazione del passato, ma l'impiego di tecniche statistiche e algoritmi di *Machine Learning* per identificare pattern nascosti e formulare stime su eventi futuri o su classificazioni ignote. Nel nostro specifico dominio applicativo, l'analisi predittiva non si traduce necessariamente in una previsione temporale (*forecasting*), ma assume la forma di una classificazione probabilistica e di una fenotipizzazione computazionale: dato un insieme di alterazioni elettroencefalografiche, qual è la combinazione di fattori che massimizza la probabilità di appartenenza a una determinata classe diagnostica?

5.1.2 Gli strumenti nativi di Intelligenza Artificiale in Power BI

Tradizionalmente, l'implementazione di modelli predittivi all'interno di ambienti di *Business Intelligence* richiedeva l'integrazione di script in linguaggi di programmazione esterni, quali R o Python. Sebbene Power BI supporti nativamente l'esecuzione di codice R e Python per la creazione di *visual* altamente personalizzati e per l'addestramento di modelli di *Machine Learning* complessi, questo approccio porta con sé alcune limitazioni architetturali: richiede competenze di programmazione specifiche, introduce latenza nei tempi di calcolo e, soprattutto, sacrifica gran parte dell'interattività dinamica (*cross-filtering*) tipica degli oggetti visivi nativi della piattaforma.

Per superare queste barriere, Microsoft ha integrato in Power BI una suite di *AI Visuals* nativi. Questi strumenti incapsulano algoritmi di *Machine Learning* allo stato dell'arte dietro un'interfaccia utente dichiarativa, eliminando la necessità di scrivere codice senza tuttavia rinunciare al rigore statistico dell'elaborazione. Tra gli strumenti nativi più rilevanti per il nostro studio spiccano:

- *Key Influencers (Fattori di Influenza Chiave)*: un motore di apprendimento supervisionato che esegue automaticamente decine di modelli di regressione logistica in parallelo per isolare le variabili categoriche o continue che guidano in modo più significativo una specifica metrica *target*.
- *Decomposition Tree (Albero di Scomposizione)*: un algoritmo di analisi della varianza *top-down* che, guidato da euristiche di *Artificial Intelligence*, esplora tutte le possibili suddivisioni spaziali dei dati per trovare i percorsi gerarchici che minimizzano o massimizzano un indicatore prestabilito.
- *Clustering Automatico*: integrabile direttamente all'interno dei grafici a dispersione, questo motore di apprendimento non supervisionato applica l'algoritmo K-Means per partizionare i *data point* in raggruppamenti omogenei nello spazio multivariato, identificando fenotipi occulti.

L'impiego esclusivo di questi *AI Visuals* nativi rappresenta una scelta progettuale precisa del nostro lavoro, orientata a dimostrare come sia oggi possibile realizzare pipeline analitiche estremamente complesse rimanendo interamente all'interno del paradigma *no-code* di Power BI, preservando, al contempo, l'assoluta interattività e fluidità del *report*.

5.2 L'analisi predittiva nel contesto della diagnostica neurologica

L'adozione di tecniche predittive e di *Machine Learning* nel settore medico, e in particolare nella neuropsichiatria, non è un mero esercizio tecnologico, ma risponde a un'urgenza clinica fondamentale: il superamento dei limiti della nosografia tradizionale. L'attuale sistema di classificazione dei disturbi mentali (DSM-5) si basa prevalentemente sull'osservazione dei sintomi clinici e su valutazioni comportamentali, un approccio che spesso raggruppa sotto la stessa etichetta diagnostica pazienti con substrati neurobiologici radicalmente diversi.

L'applicazione dell'Intelligenza Artificiale ai biomarcatori elettroencefalografici mira a fondare la cosiddetta *Psichiatria di Precisione*. In questo paradigma, l'algoritmo esplora il vasto spazio vettoriale definito dai segnali EEG alla ricerca di correlazioni latenti che l'occhio umano o la statistica descrittiva elementare non potrebbero cogliere. L'analisi predittiva consente di spostare il focus dalla domanda *"Qual è il livello medio di Asimmetria Alpha nei pazienti depressi?"* a una prospettiva molto più potente: *"Date le caratteristiche spettrali di questo paziente, qual è la probabilità che sviluppi un quadro depressivo o che appartenga a un fenotipo clinico specifico?"*.

Tuttavia, l'inserimento dell'Intelligenza Artificiale in un contesto clinico impone vincoli etici e metodologici stringenti, primo fra tutti il requisito di *interpretabilità*. Un medico non può affidare una diagnosi a un modello a scatola nera (*black-box*) che fornisce un risultato senza esplicitarne il ragionamento logico. È per questo motivo che le architetture implementate nelle nostre dashboard avanzate fanno largo uso dei principi dell'*Explainable AI* (XAI). Gli strumenti nativi di Power BI, come i Fattori di Influenza Chiave e l'Albero di Scomposizione, sono stati progettati esattamente per tradurre le ponderazioni matematiche delle reti neurali in alberi decisionali visivi e in affermazioni in linguaggio naturale, rendendo il modello totalmente trasparente, verificabile dall'operatore sanitario e difendibile dal punto di vista clinico. Nelle sezioni successive, illustreremo in dettaglio come questo framework tecnologico si traduce concretamente in due dashboard analitiche avanzate.

5.3 Dashboard 5: Vulnerabilità allo stress e controllo esecutivo

Dopo aver consolidato nel capitolo precedente una solida conoscenza dei biomarcatori spettrali globali (GSI), delle firme topologiche frontali (Beta e Gamma) e delle asimmetrie emisferiche emotive (FAA), questa sezione introduce un nuovo asse di indagine focalizzato sul rapporto tra risorse attentive e vulnerabilità allo stress. Nella Figura 5.1 riportiamo la Dashboard 5, interamente dedicata all'analisi del Rapporto Theta/Beta e alle sue implicazioni cliniche trasversali.

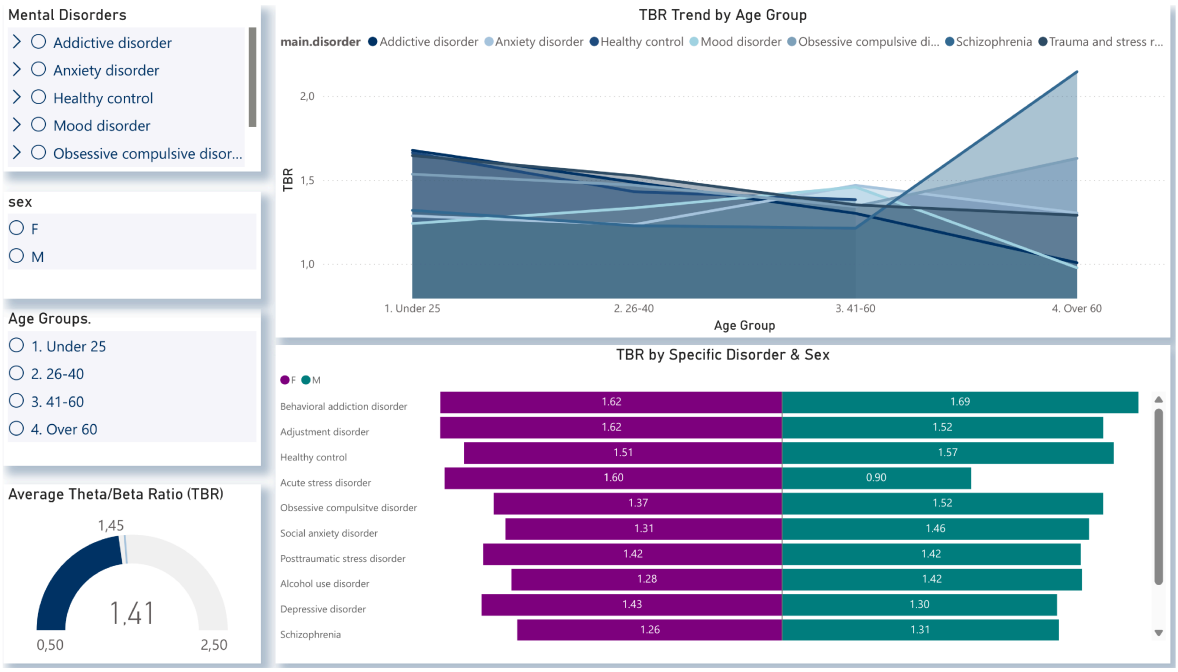


Figura 5.1: Schermata della Dashboard 5 relativa al Rapporto Theta/Beta e alla vulnerabilità allo stress esecutivo.

5.3.1 Fondamento teorico: il Rapporto Theta/Beta (TBR)

L'interpretazione clinica di questa dashboard si fonda su un biomarcatore elettroencefalografico che ha attraversato oltre due decenni di validazione scientifica e normativa: il *Theta/Beta Ratio* (TBR). L'idea sottostante è concettualmente semplice ma clinicamente potente: il rapporto tra la potenza della banda Theta (4-8 Hz), associata ai processi di introspezione, disattenzione e vagabondaggio mentale, e la potenza della banda Beta (13-30 Hz), correlata alla concentrazione attiva, al controllo inibitorio e al mantenimento del *focus* esecutivo, fornisce una misura scalare diretta dell'equilibrio attentivo del cervello [Barry *et al.*, 2003].

Un TBR equilibrato indica che la corteccia prefrontale riesce a modulare efficacemente il bilanciamento tra le risorse dedicate all'elaborazione interna (Theta) e quelle destinate al controllo attivo dell'ambiente esterno (Beta). Quando il TBR si innalza patologicamente, il rapporto si sbilancia a favore delle onde lente: il lobo frontale perde la capacità di mantenere l'inibizione comportamentale, l'attenzione sostenuta e la regolazione degli impulsi, generando quello che in letteratura viene definito come un vero e proprio "*burnout*" della corteccia prefrontale.

Dal punto di vista storico-normativo, il TBR ha raggiunto una rilevanza senza precedenti nel 2013, quando la *Food and Drug Administration* (FDA) statunitense ha autorizzato il sistema NEBA (*Neuropsychiatric EEG-Based Assessment Aid*), il primo dispositivo medico basato sull'elettroencefalografia approvato come supporto diagnostico per il Disturbo da Deficit di Attenzione e Iperattività (ADHD) [Monastra *et al.*, 1999]. Lo strumento misurava specificatamente il TBR sull'elettrodo Fz (vertice frontale), posizionato sulla linea mediana dello scalpo in corrispondenza diretta della corteccia prefrontale dorsolaterale, l'area cerebrale riconosciuta come sede primaria delle funzioni esecutive superiori. Studi meta-analitici successivi [Arns *et al.*, 2013] hanno confermato la robustezza del TBR come indicatore transdiagnostico, estendendone l'applicabilità ben oltre l'ADHD; un TBR elevato è oggi considerato un *marker* affidabile di vulnerabilità allo stress psicologico, di disfunzione della regolazione emotiva e di deficit attentivi cronici anche nei disturbi d'ansia, nei disturbi dell'umore e nei quadri post-traumatici.

Per il nostro studio, il TBR è calcolato a livello di singola osservazione (paziente per paziente) sull'elettrodo Fz, formalizzandolo nell'equazione:

$$\mathrm{TBR} = \frac{\theta_{Fz}}{\beta_{Fz}} \tag{5.3.1}$$

Un valore basso di TBR denota un cervello frontalmente attivo e capace di sostenere il controllo esecutivo; un valore criticamente elevato segnala un'iperattività delle reti di disattenzione a scapito dei circuiti di controllo, condizione che predispone il soggetto a una marcata fragilità cognitiva sotto pressione.

5.3.2 Descrizione degli oggetti visivi

L'implementazione della metrica TBR in DAX è stata realizzata mediante la funzione, preferita all'operatore aritmetico standard per la sua capacità nativa di gestire le divisioni per zero restituendo un valore nullo anziché generare un errore di calcolo. Questa scelta progettuale garantisce la resilienza dell'intera catena analitica anche in presenza di sensori malfunzionanti o di tracciati con artefatti nelle bande di interesse. Un'ulteriore accortezza tecnica ha riguardato la formattazione numerica: essendo il TBR un biomarcatore con variazioni clinicamente significative nell'ordine dei centesimi, tutti i valori sono stati forzati a due posizioni decimali nei visual, impedendo che gli arrotondamenti automatici dei grafici di *Business Intelligence* occultassero differenze diagnostiche reali.

L'architettura visiva della dashboard è articolata su tre oggetti complementari, ciascuno dei quali esamina il TBR da una prospettiva analitica distinta. Il Grafico ad Area, posizionato nella sezione superiore, mappa l'andamento del TBR medio attraverso le quattro fasce d'età definite nella fase di *feature engineering* (Under 25, 26-40, 41-60 e Over 60), stratificando le curve per macro-diagnosi. L'uso dell'area, anziché della semplice linea, come primitiva grafica, consente di percepire immediatamente il volume delle differenze tra le classi diagnostiche e di identificare le fasce temporali in cui le traiettorie divergono o convergono.

Il Grafico a Tornado (*Tornado Chart*), posizionato nella sezione inferiore, opera un confronto simmetrico *back-to-back* del TBR tra i due sessi biologici per ciascuna diagnosi specifica. Le barre maschili si estendono in una direzione e quelle femminili nella direzione opposta, creando una struttura visiva a specchio che mette in risalto le eventuali asimmetrie di genere nel controllo esecutivo.

Infine, il Grafico a Tachimetro (*Gauge Chart*) fornisce una lettura sintetica e immediata dello stato di rischio corrente. L'ago del tachimetro si posiziona sul valore medio del TBR della sottopopolazione filtrata, calibrato rispetto al *target* di normalità fissato sul valore medio dei controlli sani, pari a 1,45. La scala è stata delimitata tra un minimo di 0,50 (controllo esecutivo ottimale) e un massimo di 2,50 (deficit esecutivo severo), consentendo all'utente di valutare visivamente e con immediatezza la distanza tra il profilo corrente e il *benchmark* fisiologico.

5.3.3 Risultati e interpretazione

L'analisi del Grafico ad Area ha rivelato un andamento evolutivo del TBR che riflette fedelmente la maturazione neurofisiologica della corteccia prefrontale. Nella fascia Under 25, tutte le categorie diagnostiche presentano valori di TBR fisiologicamente elevati. Questo dato è congruente con la letteratura neuroscientifica, la quale documenta come la corteccia prefrontale sia l'ultima area cerebrale a completare la mielinizzazione e la piena maturazione sinaptica, un processo che si protrae tipicamente fino ai 25 anni di età. In questa finestra temporale, un TBR alto non è necessariamente patologico, bensì il riflesso elettrico di una corteccia frontale ancora in fase di consolidamento strutturale.

La scoperta clinicamente più significativa emerge tuttavia nella fascia 41-60 anni. Mentre la maggior parte delle categorie diagnostiche mostra una progressiva normalizzazione o stabilizzazione del TBR con l'invecchiamento, la curva relativa ai disturbi d'ansia subisce un'impennata drammatica, raggiungendo i valori più elevati dell'intero campione. Questo fenomeno, che possiamo definire come una "crisi di mezza età" elettrica, indica che la corteccia prefrontale dei pazienti ansiosi, dopo aver mantenuto un controllo esecutivo apparentemente adeguato nelle fasce giovanili e adulte, collassa funzionalmente sotto il peso cumulativo dello stress cronico. Il lobo frontale perde la capacità di sopprimere le intrusioni Theta, e il cervello scivola in uno stato di disattenzione cronica e ruminazione incontrollata.

Il Grafico a Tornado ha permesso di stratificare questa vulnerabilità per sesso biologico, svelando un *gender gap* esecutivo che attraversa trasversalmente le diagnosi. Per alcune patologie specifiche, come il disturbo da stress acuto, il divario tra maschi e femmine risulta particolarmente marcato, con il sesso femminile che esibisce valori di TBR sistematicamente più elevati. Questa asimmetria suggerisce che i meccanismi di risposta allo stress abbiano un substrato neuroelettrico sessualmente dimorfico, un'evidenza coerente con i risultati emersi dalla Dashboard 4 riguardo al dimorfismo sessuale nell'asimmetria Alpha frontale.

Il Grafico a Tachimetro, infine, restituisce per l'intera coorte un TBR medio di 1,41, un valore collocato leggermente al di sotto del *target* dei controlli sani (1,45). Questo posizionamento globale è coerente con la composizione eterogenea del campione; tuttavia, la sua vera utilità diagnostica emerge quando l'utente applica i filtri interattivi per isolare specifiche sottopopolazioni. Selezionando, ad esempio, i soli pazienti affetti da disturbo d'ansia

nella fascia 41-60, l'ago del tachimetro migra ben oltre la soglia di normalità, confermando visivamente e quantitativamente il *burnout* frontale evidenziato dal Grafico ad Area.

5.4 Dashboard 6: Integrazione di modelli di Intelligenza Artificiale

Con questa sezione, l'impianto analitico della tesi compie un salto paradigmatico: dall'analisi descrittiva e diagnostica dei biomarcatori, si approda alla classificazione predittiva e alla scoperta automatica di strutture latenti nei dati. Nella Figura 5.2 presentiamo la Dashboard 6, un pannello progettato per integrare simultaneamente tre famiglie di algoritmi di Intelligenza Artificiale — apprendimento non supervisionato, apprendimento supervisionato e analisi gerarchica della varianza — all'interno di un unico ecosistema interattivo a filtraggio incrociato.

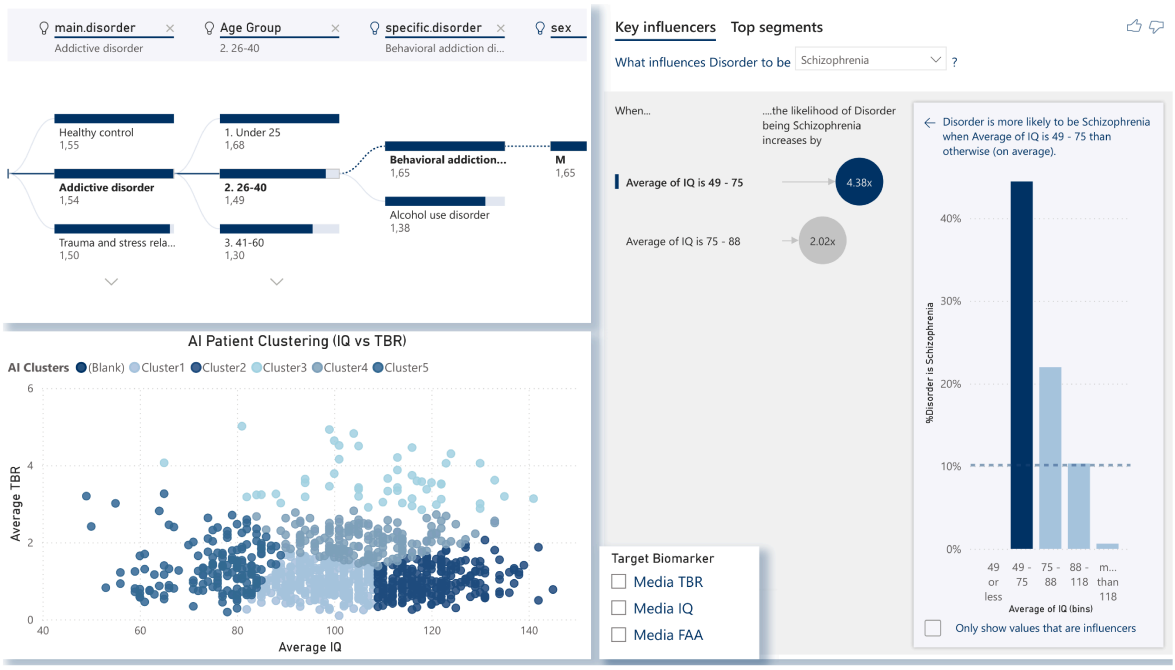


Figura 5.2: Schermata della Dashboard 6 relativa all'integrazione di modelli di Intelligenza Artificiale per la fenotipizzazione predittiva.

5.4.1 Motore di selezione dinamica e architettura della dashboard

L'elemento architetturale che distingue radicalmente questa dashboard da tutte le dashboard precedenti è la presenza di un selettore dinamico di biomarcatore *target*. Anziché vincolare l'intera infrastruttura visiva a una singola variabile dipendente predefinita, abbiamo implementato un pannello di selezione (*Slicer*) che consente all'utente di scegliere a *runtime* quale metrica debba pilotare simultaneamente tutti i modelli di Intelligenza Artificiale della schermata. Le opzioni disponibili sono tre: la *Media IQ* (quoziente intellettivo medio), la *Media TBR* (rapporto Theta/Beta medio) e la *Media FAA* (asimmetria Alpha frontale media).

Questa scelta progettuale trasforma la dashboard da un pannello di consultazione statica a un vero e proprio strumento di esplorazione scientifica parametrizzabile. Ogni commutazione del *target* ricalcola istantaneamente l'intero grafo analitico: il modello di *clustering* si riaggrega nello spazio delle *feature*, il modello predittivo ricalibra gli *Odds Ratio* e l'albero di scomposizione ricostruisce la gerarchia delle cause radice. L'interazione tra i tre motori di calcolo è ulteriormente potenziata da un'architettura di *cross-filtering* totale, nella quale la

selezione di un elemento in un qualsiasi oggetto visivo filtra e ricalcola in tempo reale tutti gli altri componenti della dashboard.

5.4.2 Apprendimento non supervisionato: Clustering K-Means

Il primo algoritmo di Intelligenza Artificiale implementato nella dashboard appartiene alla famiglia dell'apprendimento non supervisionato: il *Clustering* K-Means [MacQueen, 1967]. L'obiettivo del *clustering* non supervisionato è diametralmente opposto a quello della classificazione tradizionale: anziché assegnare ciascun paziente a una categoria diagnostica predefinita, l'algoritmo cerca autonomamente raggruppamenti naturali nei dati, ignorando completamente le etichette cliniche preesistenti. Questa indipendenza dalle diagnosi prestabilite è metodologicamente cruciale, poiché consente di identificare fenotipi neuro-cognitivi occulti che la nosografia psichiatrica tradizionale potrebbe non aver ancora codificato.

L'algoritmo K-Means opera attraverso un processo iterativo di minimizzazione della distanza euclidea intra-*cluster*. Dato un numero k di *cluster* desiderati (nel nostro caso $k = 5$), l'algoritmo inizializza k centroidi casuali nello spazio delle *feature*, assegna ogni paziente al centroide più vicino, ricalcola la posizione di ciascun centroide come baricentro dei punti assegnati e ripete il ciclo fino alla convergenza, ovvero fino a quando nessun paziente cambia *cluster* tra un'iterazione e la successiva.

Nel nostro studio, il *clustering* è stato applicato a un Grafico a Dispersione (*Scatter Plot*) che mappa il Quoziente Intellettivo sull'asse X e il TBR sull'asse Y, con i cinque *cluster* risultanti codificati cromaticamente. L'algoritmo interno di Power BI esegue la partizione direttamente sulla coppia di coordinate bidimensionali, restituendo raggruppamenti che riflettono la prossimità nello spazio cognitivo-attentivo e non la vicinanza diagnostica. L'emergere di *cluster* che attraversano trasversalmente le categorie nosografiche — contenendo, ad esempio, pazienti ansiosi e schizofrenici nella medesima aggregazione — suggerisce l'esistenza di profili neuro-cognitivi condivisi tra patologie formalmente distinte, un risultato di grande rilevanza per la psichiatria di precisione.

5.4.3 Apprendimento supervisionato: Explainable AI

Il secondo motore analitico integrato nella dashboard opera nel dominio dell'apprendimento supervisionato e dell'Intelligenza Artificiale spiegabile (*Explainable AI*, XAI) [Molnar, 2020]. A differenza del *clustering*, che opera in assenza di etichette, il modello supervisionato utilizza le diagnosi psichiatriche come variabile *target* e cerca di identificare quali combinazioni di *feature* (biomarcatori, età, sesso, istruzione) siano maggiormente predittive di una specifica condizione clinica.

L'implementazione avviene tramite l'oggetto visivo nativo di Power BI denominato Fattori di Influenza Chiave (*Key Influencers*). Al suo interno, il motore statistico esegue una Regressione Logistica [Hosmer *et al.*, 2013], un modello probabilistico che stima la probabilità di appartenenza a una classe binaria (ad esempio, "il paziente è affetto da schizofrenia" vs "il paziente non è affetto da schizofrenia") in funzione delle variabili predittive. L'output principale del modello è l'*Odds Ratio*, un moltiplicatore probabilistico che quantifica l'aumento o la diminuzione del rischio associato a ogni specifico valore di una *feature*.

L'aspetto qualificante di questo approccio risiede nella sua trasparenza algoritmica. Anziché restituire una "scatola nera" (*black-box*) con una previsione non interpretabile, il modello traduce automaticamente i coefficienti della regressione in affermazioni in linguaggio naturale direttamente leggibili e interpretabili dall'utente. Selezionando la schizofrenia come diagnosi *target*, il visual restituisce, ad esempio, che un paziente con un QI medio compreso nella fascia 49-75 ha una probabilità 4,38 volte superiore di appartenere a questa categoria

diagnostica rispetto alla popolazione generale. Analogamente, un QI nella fascia 75-88 moltiplica il rischio di un fattore 2,02. Questi risultati non solo confermano quantitativamente il noto deficit cognitivo associato ai disturbi psicotici, ma forniscono soglie numeriche precise e clinicamente utilizzabili per la stratificazione del rischio.

5.4.4 Analisi della varianza: Decomposition Tree

Il terzo e ultimo algoritmo integrato nella dashboard è l'Albero di Scomposizione (*Decomposition Tree*), uno strumento di analisi gerarchica *top-down* progettato per la *Root Cause Analysis* [Microsoft, 2024]. L'obiettivo di questo visual è identificare le specifiche combinazioni di attributi categorici che massimizzano o minimizzano una metrica continua di interesse, scomponendo ricorsivamente la varianza totale in contributi sempre più granulari.

L'albero opera con un meccanismo di ramificazione *top-down* assistito dall'Intelligenza Artificiale (*Split AI*): ad ogni livello, l'algoritmo valuta tutte le variabili categoriche disponibili (macro-diagnosi, fascia d'età, diagnosi specifica, sesso) e seleziona automaticamente la dimensione di *split* che produce la massima separazione nella metrica *target*. L'utente può anche guidare manualmente la discesa nell'albero, scegliendo, ad ogni nodo, quale variabile esplorare, creando un dialogo interattivo tra intuizione clinica e guida algoritmica.

L'albero visualizzato nella Dashboard 6, partendo dal nodo radice che rappresenta il valore medio del biomarcatore selezionato per l'intera coorte, effettua una discesa progressiva attraverso le dimensioni diagnostiche, anagrafiche e biologiche. Ad ogni ramificazione, i valori numerici associati ai nodi figli espongono le sotto-medie filtrate, permettendo di isolare con precisione chirurgica le specifiche combinazioni fenotipiche responsabili delle alterazioni di segnale più critiche.

5.4.5 Il loop analitico: integrazione e validazione incrociata

L'eccellenza metodologica di questa dashboard risiede non nei singoli modelli, ciascuno dei quali ha limiti intrinseci se considerato isolatamente, bensì nella loro integrazione sinergica attraverso il meccanismo di *cross-filtering* totale. L'architettura di filtraggio incrociato trasforma tre algoritmi indipendenti in un unico ciclo analitico iterativo capace di generare scoperte che nessuno dei componenti avrebbe potuto produrre autonomamente.

Il ciclo opera come segue: selezionando un *cluster* K-Means nel Grafico a Dispersione, l'utente isola una sottopopolazione di pazienti accomunati dalla vicinanza nello spazio cognitivo-attentivo, indipendentemente dalla loro diagnosi formale. Questa selezione filtra simultaneamente l'Albero di Scomposizione, che si contrae per mostrare esclusivamente la struttura gerarchica interna a quel *cluster*, e i Fattori di Influenza Chiave, che ricalcolano gli *Odds Ratio* della regressione logistica purificandoli dal rumore delle altre sottopopolazioni. L'effetto netto è una decontaminazione statistica progressiva: ogni livello di filtraggio rimuove varianza confondente, affinando la precisione predittiva e facendo emergere regolarità biomediche che nell'aggregato completo sarebbero state diluite o mascherate.

Viceversa, il ciclo può essere percorso in direzione opposta: partendo da una regola decisionale identificata dalla regressione logistica (ad esempio, "il rischio di schizofrenia aumenta di 4,38 volte per QI nella fascia 49-75"), l'utente può selezionare questa fascia di rischio e osservare come i *cluster* K-Means si ridistribuiscono, verificando se i pazienti ad alto rischio convergano in uno o più raggruppamenti specifici o siano dispersi trasversalmente.

Questo dialogo bidirezionale tra apprendimento supervisionato e non supervisionato trasforma un database statico in un motore di scoperta iterativa. L'infrastruttura non si limita a confermare ipotesi cliniche preesistenti, ma genera attivamente nuove domande di ricerca, identificando sottogruppi di pazienti con profili neuro-cognitivi anomali che meritano un'indagine clinica dedicata. L'integrazione di tre paradigmi algoritmici distinti

in un unico pannello interattivo rappresenta, a nostro giudizio, il contributo metodologico più significativo dell'intero progetto di tirocinio, dimostrando come gli strumenti di *Business Intelligence* possano essere efficacemente impiegati non solo per la reportistica descrittiva, ma anche come veri e propri ambienti di ricerca esplorativa in ambito biomedico.

CAPITOLO 6

Discussione

Il presente capitolo costituisce il momento di sintesi critica dell'intero percorso di ricerca. Nella prima parte, i risultati ottenuti attraverso le sei dashboard analitiche vengono messi a confronto con le evidenze della letteratura neuroscientifica e psichiatrica, al fine di valutarne la coerenza e il significato clinico. Nella seconda parte, il lavoro viene esaminato alla luce dei suoi punti di forza strutturali e delle sue limitazioni metodologiche.

6.1 Discussione dei risultati

Il progetto ha prodotto un insieme articolato di *insight* neurofisiologici, derivati dall'incrocio di biomarcatori EEG con le etichette diagnostiche psichiatriche di 941 pazienti. In questa sezione, i principali risultati vengono esaminati criticamente e confrontati con la letteratura scientifica di riferimento, al fine di valutarne la robustezza e il valore scientifico.

6.1.1 Il Global Slowing Index come discriminante patologico trasversale

La Dashboard 2 ha quantificato il *Global Slowing Index* (GSI) su base demografica e diagnostica, restituendo un valore medio per l'intera coorte di 1,42. Questo valore, di per sé prossimo alla neutralità, ha acquistato tutto il suo potere discriminante quando analizzato per sottogruppi: patologie gravi, come la schizofrenia cronica e le demenze, presentano valori di GSI marcatamente elevati, con le componenti Delta e Theta che arrivano a occupare oltre il 60% dell'energia cerebrale totale. Questo risultato è in piena coerenza con la letteratura neuroscientifica: già Klimesch [Klimesch, 1999] aveva documentato il progressivo aumento delle potenze lente come correlato elettrico della perdita di sincronizzazione talamocorticale, un meccanismo fisiopatologico centrale nella neurodegenerazione. L'utilizzo del *Box Plot* ha permesso un'osservazione ulteriore di grande rilevanza: l'elevatissima varianza intra-diagnostica della schizofrenia (visualizzata dall'allungamento verticale delle "scatole") suggerisce che questa etichetta nosografica racchiuda in realtà fenotipi elettrici eterogenei, alcuni dei quali condividono caratteristiche con la schizofrenia in stadio precoce e altri con forme conclamate e croniche. Questa frammentazione interna all'etichetta diagnostica rappresenta uno degli argomenti più forti a favore di una psichiatria di precisione basata sui biomarcatori.

6.1.2 Efficienza neurale e sincronizzazione Gamma: la conferma del gradiente

La Dashboard 3 ha testato empiricamente l'ipotesi dell'efficienza neurale [Neubauer e Fink, 2009], rilevando una correlazione positiva tra il quoziente intellettivo e la potenza Beta frontale nei soggetti sani. Nei controlli, l'aumento del QI è accompagnato da un incremento proporzionale dell'attivazione nelle bande rapide prefrontali, una conferma visiva del concetto di "routing" selettivo delle risorse neurali. Altrettanto significativo è il risultato relativo alla banda Gamma: il Grafico a Violino ha mostrato una distribuzione bimodale per i pazienti schizofrenici, con una coda destra anomala che indica la presenza di un sottogruppo con attività Gamma paradossalmente iperattiva. Questo pattern non è privo di precedenti in letteratura: alcuni autori [Uhlhaas e Singer, 2010] hanno descritto un'iperattivazione Gamma nelle fasi acute della psicosi come conseguenza di un fallimento del controllo inibitorio inter-neuronale, un meccanismo distinto dal più classico deficit di sincronizzazione Gamma documentato nelle forme croniche. La Dashboard 3 ha, dunque, catturato, attraverso la morfologia della KDE, una distinzione clinica che i soli valori medi non avrebbero mai potuto rivelare.

6.1.3 L'asimmetria Alpha Frontale e il dimorfismo sessuale nell'OCD

Il risultato clinicamente più inatteso e originale dell'intero studio è emerso dalla Dashboard 4, nell'analisi del Disturbo Ossessivo-Compulsivo (OCD) stratificato per sesso biologico. I pazienti maschili con OCD presentano un indice FAA positivo (dominanza sinistra, sistema BAS attivo), mentre le pazienti femminili con la medesima diagnosi mostrano un FAA fortemente negativo (dominanza destra, sistema BIS iperattivo). Questa divergenza di polarità all'interno della stessa categoria diagnostica è di notevole interesse scientifico. L'asimmetria destrorsa nelle donne con OCD potrebbe riflettere una maggiore predisposizione alla componente ansiosa ruminativa e inibitoria del disturbo, che si sovrappone al nucleo ossessivo-compulsivo puro, configurando un fenotipo clinico misto e più complesso. Questa ipotesi è coerente con i dati epidemiologici che documentano come le donne con OCD tendano a presentare una maggiore prevalenza di ossessioni di contaminazione e di rituali di controllo rispetto agli uomini, comportamenti intrinsecamente guidati dalla logica dell'evitamento (BIS). Il risultato suggerisce che la FAA possa fungere da indicatore di rischio per l'identificazione precoce del sottotipo ansioso-inibitorio dell'OCD, con ricadute potenziali nella personalizzazione del percorso terapeutico.

6.1.4 Il Theta/Beta Ratio e la "crisi di mezza età" corticale

La Dashboard 5 ha rivelato un pattern evolutivo del TBR particolarmente eloquente: la curva relativa ai disturbi d'ansia, relativamente contenuta nelle fasce giovanili, registra un'impennata drammatica nella fascia 41–60 anni, raggiungendo i valori più elevati dell'intero campione. Questo fenomeno, interpretabile come un collasso tardivo del controllo esecutivo frontale sotto il peso dello stress cronico accumulato, ha una spiegazione neurobiologica precisa. È noto che la corteccia prefrontale, in condizioni di stress cronico protratto, subisce una riduzione progressiva della densità dendritica e una retrazione delle spine sinaptiche, con conseguente diminuzione della capacità di produrre il ritmo Beta ad alta frequenza necessario per il controllo inibitorio [McEwen e Morrison, 2013]. Quando questa riserva strutturale si esaurisce nella mezza età, il TBR si sbilanicia irreversibilmente a favore del Theta, generando lo stato di disattenzione cronica e ruminazione incontrollata osservato. Questo risultato ha implicazioni cliniche dirette; esso suggerisce l'esistenza di una finestra critica di vulnerabilità nella quinta decade di vita per i pazienti ansiosi, durante la quale

un intervento preventivo mirato (tecniche di regolazione del sistema nervoso autonomo, *mindfulness* o neurofeedback) potrebbe ritardare o mitigare il deterioramento esecutivo.

6.1.5 Il Clustering K-Means e la fenotipizzazione trasversale

La scoperta metodologicamente più rilevante della Dashboard 6 riguarda il comportamento del *clustering* K-Means: i cinque raggruppamenti identificati dall'algoritmo nel piano cartesiano QI/TBR attraversano trasversalmente le categorie diagnostiche tradizionali. In altri termini, alcuni *cluster* raggrupano pazienti con diagnosi formalmente distinte (ad esempio, un paziente ansioso e un paziente schizofrenico a basso QI e alto TBR) all'interno del medesimo profilo neuro-cognitivo. Questo risultato non è una disfunzione algoritmica, bensì una conferma computazionale di un'intuizione crescente in psichiatria: i disturbi mentali non sono entità discrete e nettamente separate, ma si sovrappongono e si intersecano in uno spazio dimensionale continuo di tratti e biomarcatori. Il *clustering* EEG non supervisonato ha, dunque, prodotto una tassonomia alternativa, fondata sull'elettrofisiologia e non sulla fenomenologia comportamentale, che potrebbe integrare e arricchire i sistemi diagnostici tradizionali come il DSM-5.

6.2 Punti di forza e limitazioni dello studio

6.2.1 Punti di forza

Il presente studio presenta diversi elementi di originalità e robustezza. In primo luogo, l'ampiezza del campione analizzato (941 soggetti) e la straordinaria eterogeneità delle categorie diagnostiche rappresentate, che spaziano dai disturbi d'umore alle dipendenze, dai disturbi d'ansia ai controlli sani, conferiscono ai risultati una buona generalizzabilità e consentono analisi comparative cross-diagnostiche di elevato valore scientifico.

In secondo luogo, l'approccio metodologico adottato — la costruzione di un intero ecosistema analitico integrato all'interno di Power BI, senza ricorrere a linguaggi di programmazione esterni — dimostra concretamente la maturità e la potenza delle piattaforme di *Business Intelligence* moderna nel contesto della ricerca biomedica. Questo approccio *no-code* abbassa significativamente la barriera tecnica di accesso all'analisi avanzata, rendendo le dashboard potenzialmente utilizzabili anche da personale clinico privo di competenze informatiche avanzate.

In terzo luogo, l'integrazione di biomarcatori validati dalla letteratura scientifica internazionale (GSI, FAA, TBR) garantisce che le metriche elaborate non siano costrutti arbitrari, ma misurazioni con un solido fondamento teorico e sperimentale. Infine, l'adozione dell'approccio XAI (*Explainable AI*) nella Dashboard 6 garantisce la trasparenza algoritmica, un requisito non negoziabile per qualsiasi applicazione clinica di strumenti di Intelligenza Artificiale.

6.2.2 Limitazioni

Il lavoro presenta, tuttavia, alcune limitazioni metodologiche che è doveroso riconoscere in modo esplicito. La più rilevante riguarda la natura trasversale (*cross-sectional*) del dataset: ciascuna osservazione rappresenta una fotografia istantanea dell'attività elettrica del paziente in un unico momento, senza alcuna informazione sull'evoluzione temporale della patologia, sulle terapie farmacologiche in corso o sull'anamnesi clinica. Questa assenza di dati longitudinali impedisce qualsiasi inferenza causale e limita le conclusioni a livello correlazionale.

Una seconda limitazione riguarda il *class imbalance*: alcune categorie diagnostiche (disturbi dell'umore, schizofrenia, disturbi d'ansia) sono notevolmente più rappresentate di altre (come

le dipendenze o i disturbi correlati a trauma), un'asimmetria che potrebbe distorcere i modelli predittivi supervisionati, i quali tendono a privilegiare le classi maggioritarie. Sebbene i modelli nativi di Power BI non consentano un controllo esplicito di questo fenomeno tramite tecniche come il *SMOTE* o la pesatura delle classi, la consapevolezza di questo limite è essenziale per una corretta interpretazione degli *Odds Ratio* prodotti dalla regressione logistica.

Infine, un limite di carattere tecnico riguarda la modalità di calcolo del TBR e del GSI: entrambi i biomarcatori sono stati derivati dai valori di potenza spettrale medi presenti nel dataset, che rappresentano aggregati pre-elaborati e non i segnali EEG grezzi originali. La pipeline di analisi completa del segnale grezzo (con artefact rejection, filtraggio e trasformata di Fourier) avrebbe potuto introdurre ulteriori sfumature di precisione non catturabili a partire dai dati tabulari.

Conclusioni

Il presente lavoro di tesi ha intrapreso un percorso di ricerca all'intersezione tra la neurofisiologia clinica, la Data Science e la Business Intelligence. L'obiettivo principale è stato quello di dimostrare come gli strumenti di visualizzazione e analisi dei dati tipici del mondo aziendale potessero essere ricondotti con successo nell'ambito della ricerca biomedica, superando la logica della mera reportistica descrittiva per approdare a un paradigma di scoperta scientifica esplorativa. A tal fine, abbiamo progettato e sviluppato un ecosistema analitico integrato all'interno di Microsoft Power BI, sfruttando un approccio interamente no-code che rende lo strumento accessibile anche a personale medico privo di competenze informatiche avanzate. Le sei dashboard interattive realizzate hanno consentito di estrarre, elaborare e interpretare biomarcatori neurofisiologici complessi a partire da un dataset di 941 pazienti affetti da disturbi psichiatrici eterogenei. Abbiamo quantificato l'impatto del Global Slowing Index nella discriminazione delle patologie, confermato l'ipotesi dell'efficienza neurale in correlazione con la potenza Beta e rilevato pattern originali, quali la divergenza di polarità dell'asimmetria Alpha frontale tra sessi all'interno del Disturbo Ossessivo-Compulsivo. L'integrazione nativa di modelli di Intelligenza Artificiale, come il clustering K-Means e la Regressione Logistica, ci ha inoltre permesso di identificare fenotipi neuro-cognitivi che attraversano trasversalmente le categorie diagnostiche classiche, suggerendo una nuova tassonomia basata sulla firma elettrica del cervello. I risultati conseguiti aprono prospettive di sviluppo futuro particolarmente ampie, sia sul versante algoritmico che su quello dell'applicazione clinica. Dal punto di vista tecnico, il naturale passo successivo consiste nell'addestramento di architetture di Deep Learning, come le reti neurali ricorrenti o i modelli Transformer, direttamente sul segnale elettroencefalografico grezzo nel dominio del tempo, al fine di catturare dinamiche ritmiche non rilevabili attraverso i soli valori di potenza spettrale media. Parallelamente, si prevede l'integrazione di flussi informativi multimodali, combinando i tracciati EEG con i dati di neuroimaging funzionale e test psicometrici per una stratificazione del rischio biologico ancora più accurata. Sul fronte prettamente applicativo, il modello predittivo sviluppato necessiterà di una rigorosa validazione prospettica in ambiente ospedaliero su dati raccolti ex novo, passaggio imprescindibile per una sua eventuale approvazione come dispositivo software di supporto decisionale. Infine, l'integrazione di tali sistemi analitici con le moderne infrastrutture IoT e dispositivi indossabili promette di trasformare queste dashboard in cruscotti di monitoraggio longitudinale, capaci di tracciare in tempo reale la risposta del paziente alla terapia e anticipare le ricadute sintomatiche, tracciando così la rotta verso una psichiatria sempre più personalizzata, preventiva e di precisione.

Bibliografia

ABOU JAOUDE, M., SUN, H., PELLERIN, K. A., BHATT, D. e WESTOVER, M. B. (2020), «Expert-level automated sleep staging of long-term scalp electroencephalography recordings using deep learning», *Sleep*, vol. 43 (11). (Cited at page 6)

ARNS, M., CONNERS, C. K. e KRAEMER, H. C. (2013), «A Decade of EEG Theta/Beta Ratio Research in ADHD: A Meta-Analysis», *Journal of Attention Disorders*, vol. 17 (5), p. 374–383. (Cited at page 53)

BANERJEE, A., BANDYOPADHYAY, T. e ACHARYA, P. (2013), «Data Analytics: Hyped Up Aspirations or True Potential?», *Vikalpa: The Journal for Decision Makers*, vol. 38 (4), p. 1–12. (Cited at pages 8, 9 e 10)

BARRY, R. J., CLARKE, A. R. e JOHNSTONE, S. J. (2003), «A Review of Electrophysiology in Attention-Deficit/Hyperactivity Disorder: I. Qualitative and Quantitative Electroencephalography», *Clinical Neurophysiology*, vol. 114 (2), p. 171–183. (Cited at page 53)

BZDOK, D. e MEYER-LINDENBERG, A. (2018), «Machine Learning for Precision Psychiatry: Opportunities and Challenges», *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 3 (3), p. 223–230. (Cited at pages 7 e 8)

CODD, E. F. (1970), «A Relational Model of Data for Large Shared Data Banks», *Communications of the ACM*, vol. 13 (6), p. 377–387. (Cited at page 3)

DAVENPORT, T. H. e HARRIS, J. G. (2007), *Competing on Analytics: The New Science of Winning*, Harvard Business Press. (Cited at pages 7 e 8)

DAVIDSON, R. J. (1992), «Emotion and affective style: Hemispheric substrates», *Psychological science*, vol. 3 (1), p. 39–43. (Cited at page 47)

DELEN, D. e DEMIRKAN, H. (2013), «Data, information and analytics as services», *Decision Support Systems*, vol. 55 (1), p. 218–225. (Cited at pages 9, 10 e 11)

DOMO (2020), «Data Never Sleeps 8.0», Infografica, disponibile online: `https://www.domo.com/learn/infographic/data-never-sleeps-8`. (Cited at page 4)

ERL, T., KHATTAK, W. e BUHLER, P. (2016), *Big Data Fundamentals: Concepts, Drivers & Techniques*, Prentice Hall. (Cited at pages 11, 12, 13, 14, 15, 16, 17 e 18)

GANDOMI, A. e HAIDER, M. (2015), «Beyond the hype: Big data concepts, methods, and analytics», *International Journal of Information Management*, vol. 35 (2), p. 137–144. (Cited at pages 5 e 7)

GARTNER (2012), «Big Data», Gartner IT Glossary, disponibile online: `https://www.gartner.com/en/information-technology/glossary/big-data`. (Cited at page 3)

GRAY, J. A. (1981), *A critique of Eysenck's theory of personality*, Springer. (Cited at page 48)

HEY, T., TANSLEY, S. e TOLLE, K. (2009), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research. (Cited at page 4)

HILBERT, M. e LÓPEZ, P. (2011), «The World's Technological Capacity to Store, Communicate, and Compute Information», *Science*, vol. 332 (6025), p. 60–65. (Cited at page 6)

HOSMER, D. W., LEMESHOW, S. e STURDIVANT, R. X. (2013), *Applied Logistic Regression*, John Wiley & Sons, Hoboken, NJ, 3rd ed. (Cited at page 56)

ISLAM, M. K., RASTEGARNIA, A. e YANG, Z. (2016), «Methods for artifact detection and removal from scalp EEG: A review», *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 46 (4-5), p. 287–305. (Cited at page 7)

KIMBALL, R. e ROSS, M. (2013), *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, 3rd ed. (Cited at page 23)

KLIMESCH, W. (1999), «EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis», *Brain research reviews*, vol. 29 (2-3), p. 169–195. (Cited at page 59)

LANEY, D. (2001), «3D Data Management: Controlling Data Volume, Velocity and Variety», META Group Research Note. (Cited at page 5)

MACQUEEN, J. (1967), «Some Methods for Classification and Analysis of Multivariate Observations», in «Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics», p. 281–297, University of California Press, Berkeley, CA. (Cited at page 56)

MANYIKA, J., CHUI, M., BROWN, B., BUGHIN, J., DOBBS, R., ROXBURGH, C. e BYERS, A. H. (2011), «Big data: The next frontier for innovation, competition, and productivity», Rap. tecn., McKinsey Global Institute. (Cited at page 4)

MCEWEN, B. S. e MORRISON, J. H. (2013), «The brain on stress: vulnerability and plasticity of the prefrontal cortex over the life course», *Neuron*, vol. 79 (1), p. 16–29. (Cited at page 60)

MICROSOFT (2024), «Create and View Decomposition Tree Visuals in Power BI», `https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-decomposition-tree`, accessed: 2024. (Cited at page 57)

MICROSOFT CORPORATION (2024), «What is Power BI?», URL `https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview`, accessed: 2025-06-23. (Cited at pages 19 e 21)

MOLNAR, C. (2020), *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, Christoph Molnar, URL `https://christophm.github.io/interpretable-ml-book/`. (Cited at page 56)

MONASTRA, V. J., LUBAR, J. F., LINDEN, M., VANDEUSEN, P., GREEN, G., WING, W., PHILLIPS, A. e FENGER, T. N. (1999), «Assessing Attention Deficit Hyperactivity Disorder via Quantitative Electroencephalography: An Initial Validation Study», *Neuropsychology*, vol. 13 (3), p. 424–433. (Cited at page 53)

NEUBAUER, A. C. e FINK, A. (2009), «Intelligence and neural efficiency», *Neuroscience & Biobehavioral Reviews*, vol. 33 (7), p. 1004–1023. (Cited at pages 45 e 60)

POWER, D. J. (2002), *Decision Support Systems: Concepts and Resources for Managers*, Quorum Books. (Cited at page 8)

PRICHEP, L. S. e JOHN, E. R. (1992), «QEEG profiles of psychiatric disorders», *Brain topography*, vol. 4 (4), p. 249–257. (Cited at pages 28 e 30)

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. (Cited at page 8)

RAGHUPATHI, W. e RAGHUPATHI, V. (2014), «Big data analytics in healthcare: promise and potential», *Health Information Science and Systems*, vol. 2 (1), p. 3. (Cited at pages 4, 6 e 8)

RESEARCH, G. (2026), «Magic Quadrant for Analytics and Business Intelligence Platforms», Rap. tecn., Gartner, Inc., URL `https://www.gartner.com/en/research/magic-quadrant`, microsoft positioned as Leader for the 19th consecutive year. (Cited at page 20)

RUSSO, M. e FERRARI, A. (2019), *The Definitive Guide to DAX: Business Intelligence with Microsoft Power BI, SQL Server Analysis Services, and Excel*, Microsoft Press, 2nd ed. (Cited at pages 19, 20, 22, 26 e 27)

UHLHAAS, P. J. e SINGER, W. (2010), «Abnormal neural oscillations and synchrony in schizophrenia», *Nature reviews neuroscience*, vol. 11 (2), p. 100–113. (Cited at page 60)

WORLD HEALTH ORGANIZATION (2019), «The WHO special initiative for mental health (2019-2023): Universal Health Coverage for Mental Health», WHO Report. (Cited at page 4)

Sitografia

- Microsoft Power BI Documentation – learn.microsoft.com/en-us/power-bi
- Microsoft Power Query Documentation – learn.microsoft.com/en-us/power-query
- Microsoft DAX Reference – learn.microsoft.com/en-us/dax
- Power BI Official Blog – powerbi.microsoft.com/en-us/blog
- DAX Guide (SQLBI) – dax.guide
- SQLBI (Alberto Ferrari & Marco Russo) – www.sqlbi.com
- Kaggle: EEG Psychiatric Disorders Dataset – www.kaggle.com/datasets/shashwatwork/eeg-psychiatric-disorders-dataset
- Kaggle Data Science Community – www.kaggle.com
- OMS (Organizzazione Mondiale della Sanità): Mental Health – www.who.int/health-topics/mental-health
- PubMed (National Institutes of Health) – pubmed.ncbi.nlm.nih.gov
- IEEE Xplore Digital Library – ieeexplore.ieee.org
- Gartner Magic Quadrant Research – www.gartner.com
- TechTarget: Data Management & Analytics – www.techtarget.com/searchdatamanagement
- Scikit-Learn: Machine Learning in Python – scikit-learn.org
- Python Official Documentation – docs.python.org
- Towards Data Science – towardsdatascience.com
- KDnuggets: Data Science, Machine Learning, AI – www.kdnuggets.com
- World Psychiatric Association (WPA) – www.wpanet.org
- FDA (Food and Drug Administration): Press Announcements – www.fda.gov/news-events
- IBM Data and AI Analytics – www.ibm.com/analytics

Ringraziamenti

Desidero rivolgere il mio primo ringraziamento alla mia famiglia, per avermi sempre sostenuto durante questo intero percorso universitario, dandomi la possibilità e la forza di portarlo a termine con serenità. Un ringraziamento speciale e di cuore va a mia sorella, Cynthia, che è stata per me di un aiuto insostituibile, sostenendomi sin da quando eravamo piccoli.

Un sincero ringraziamento va al Prof. Domenico Ursino, relatore di questa tesi, per avermi guidato nello sviluppo del progetto con estrema disponibilità, offrendomi spunti di riflessione fondamentali e un prezioso supporto continuo.

Infine, ringrazio tutti i miei amici con i quali ho condiviso le fatiche e le soddisfazioni di questi anni di studio.