UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Corso di Laurea Triennale in Informatica

TESI DI LAUREA

Comorbidity Analysis and Disease Evolution: A Platform for Visualizing Patterns through Social Network Analysis

RELATORI

Prof. Delfina Malandrino

Prof. Rocco Zaccagnino

Università degli Studi di Salerno

CANDIDATO

Gianfranco Barba

Matricola: 0512114635

Anno Accademico 2024-2025

”Success is not final, failure is not fatal:
It is the courage to continue that counts."
– Winston Churchill

Abstract

Questa tesi affronta lo studio della comorbidità attraverso l'analisi dei dati clinici, con l'obiettivo di comprendere meglio le relazioni tra le malattie e fornire supporto decisionale ai professionisti sanitari. Il lavoro si è sviluppato in diverse fasi: inizialmente è stato condotto uno studio approfondito sulla comorbidità e sui pattern ricorrenti nei dati sanitari strutturati.

Successivamente, è stata sviluppata ComorGraph, una piattaforma di Visual Analytics e Social Network Analysis (SNA) integrata nell'ecosistema MedMiner, che permette di rappresentare le comorbidità sotto forma di grafo, identificando malattie chiave e le loro interconnessioni. La piattaforma utilizza metriche grafiche per analizzare le relazioni tra le patologie.

Infine, lo studio si arricchisce di un modulo di Intelligenza Artificiale basato su Graph Neural Network (GNN), che, pur non essendo parte integrante di ComorGraph, consente di effettuare previsioni sull'insorgenza di nuove malattie nei pazienti, offrendo uno strumento avanzato per la gestione delle comorbidità.

Questa tesi è stata realizzata in

EDGE
RESEARCH LAB

Indice

Elenco delle figure

Elenco delle tabelle

CAPITOLO 1

Introduzione

1.1 Contesto Applicativo

Lo studio della comorbidità, ovvero la coesistenza di più malattie in un paziente, è diventato un argomento cruciale nella medicina moderna. Comprendere come le patologie si influenzano reciprocamente è fondamentale per migliorare la gestione clinica dei pazienti con condizioni complesse. La crescita esponenziale dei dati sanitari e l'interconnessione tra pazienti e malattie richiede l'uso di strumenti innovativi per analizzare queste relazioni.

La rappresentazione dei dati clinici sotto forma di grafo offre una visione più chiara delle correlazioni tra le malattie. Un grafo permette di modellare pazienti e patologie come nodi, con le connessioni tra malattie rappresentate da archi, facilitando l'analisi della comorbidità tramite metriche di Social Network Analysis (SNA). Tuttavia, gli approcci tradizionali non considerano la dimensione temporale, rendendo difficile prevedere l'evoluzione clinica di un paziente.

Per affrontare queste sfide, è stata sviluppata **ComorGraph**, una piattaforma integrata nell'ecosistema **MedMiner** (progettato per supportare medici e specialisti nell'analisi di casi clinici complessi). **ComorGraph** sfrutta la rappresentazione grafica dei dati clinici per analizzare i pattern di correlazione tra malattie, consentendo

una maggiore precisione nelle prescrizioni terapeutiche e supportando la gestione dell'insorgenza di nuove patologie strettamente correlate a quelle già presenti nel quadro clinico del paziente.

A complemento della piattaforma, è stato condotto uno studio sull'applicazione del **Machine Learning** per facilitare l'analisi di questi pattern. Data la natura grafica della rappresentazione, è stato sviluppato un modello basato su **Graph Neural Networks (GNN)**, in grado di rilevare pattern complessi e meno evidenti, e di prevedere la probabilità che un paziente sviluppi determinate patologie in base al proprio quadro clinico. Questo approccio consente di delineare terapie personalizzate mirate non solo alla cura, ma anche alla prevenzione delle malattie con maggiore probabilità di insorgenza.

1.2 Motivazioni e Obiettivi

La gestione delle comorbidità, specialmente in pazienti con più patologie croniche, richiede strumenti avanzati per identificare le relazioni tra malattie e prevedere possibili sviluppi clinici. Gli approcci tradizionali si concentrano su modelli statici, spesso non sufficienti a cogliere l'interconnessione dinamica tra le patologie né a considerare la storia clinica del paziente nel tempo. Questa limitazione comporta difficoltà per i professionisti sanitari nel prendere decisioni preventive e nel gestire le comorbidità in modo efficiente.

Per rispondere a queste sfide, è stata sviluppata **ComorGraph**, una piattaforma innovativa che combina la rappresentazione grafica dei dati sanitari con algoritmi di intelligenza artificiale per fornire previsioni cliniche personalizzate.

◎ **Our Goal.** Sviluppare una piattaforma innovativa per la **Visual Analytics** di dati clinici complessi, fornendo uno strumento avanzato per l'analisi delle comorbidità. Integrando metriche di **Social Network Analysis (SNA)**, la piattaforma consentirà di identificare pattern tra le malattie e supportare i medici nell'individuazione di patologie critiche, offrendo un contributo essenziale alla gestione dei casi clinici complessi.

1.3 Risultati Ottenuti

La sperimentazione condotta con **ComorGraph** ha prodotto risultati significativi nell'analisi della comorbidità e nella predizione delle malattie. La piattaforma è stata testata utilizzando un dataset di dati clinici reali, permettendo di identificare con precisione le relazioni tra le malattie e di evidenziare patologie particolarmente connesse tra loro.

Grazie all'applicazione delle metriche di **Social Network Analysis (SNA)**, è stato possibile individuare nodi centrali nel grafo delle malattie, fornendo nuove intuizioni sui collegamenti tra le patologie e sul loro impatto nei casi clinici. Inoltre, il modulo di intelligenza artificiale basato su una **Graph Neural Network (GNN)** ha dimostrato una capacità predittiva promettente, consentendo di stimare con buona accuratezza la probabilità di comparsa di una malattia specifica in un paziente nel tempo.

Questi risultati aprono nuove opportunità nell'uso dei grafi e dell'IA per migliorare la comprensione delle comorbidità e supportare i professionisti sanitari nella gestione dei pazienti con condizioni complesse.

1.4 Struttura della Tesi

Questa tesi è suddivisa in cinque capitoli, ognuno dei quali approfondisce aspetti specifici del lavoro di ricerca e sviluppo svolto.

- **Capitolo 2: Background e Stato dell'Arte** – Questo capitolo introduce i concetti chiave e le metodologie esistenti per lo studio della comorbidità tramite grafi, fornendo una panoramica delle tecniche di **Social Network Analysis (SNA)**, **Visual Analytics (VA)** e delle reti neurali grafiche (**GNN**) applicate al contesto medico.

- **Capitolo 3: Metodologia e Implementazione** – Viene descritto il processo di sviluppo della piattaforma **ComorGraph**, con dettagli sull'architettura del sistema e l'introduzione del modello predittivo. Il codice sorgente della piattaforma, insieme alla documentazione dettagliata, è reso disponibile pub-

blicamente sul repository GitHub al seguente link: `https://github.com/gianfrancobarba/MedMiner`.

- **Capitolo 4: Pre-Processing e Bilanciamento del Dataset Clinico** – Questo capitolo descrive in maniera chiara e approfondita il processo che ha portato alla creazione di un dataset bilanciato e coerente con le esigenze della piattaforma.
- **Capitolo 5: Social Network Analysis (SNA) e Visual Analytics (VA) per l'Analisi della Comorbidità** - Questo capitolo approfondisce le metriche di Social Network Analysis trattate sulla piattaforma e mostra come vengono visualizzate.
- **Capitolo 6: Conclusioni e Sviluppi Futuri** – Viene discusso l'impatto della piattaforma **ComorGraph** e i potenziali sviluppi futuri, con particolare attenzione alle possibili estensioni del sistema e alle sfide future nel campo della medicina predittiva.

CAPITOLO 2

Background e Stato dell'Arte

2.1 La comorbidità: una sfida clinica complessa

La comorbidità, definita come la coesistenza di due o più malattie croniche in uno stesso individuo, rappresenta una delle maggiori sfide per la medicina moderna. Questo fenomeno non solo complica la gestione clinica dei pazienti, ma impatta anche significativamente sulla qualità della vita e sulle risorse sanitarie. La presenza di più condizioni patologiche contemporanee aumenta il rischio di esiti clinici negativi, quali una maggiore mortalità, un peggioramento della qualità della vita e un aumento dei costi sanitari globali [1].

La distinzione tra comorbidità e multimorbidità, termini spesso usati in modo intercambiabile, è rilevante dal punto di vista clinico. Il termine comorbidità solitamente si riferisce alla presenza di altre malattie in aggiunta a una condizione principale (ad esempio, il diabete associato a ipertensione), mentre la multimorbidità implica l'assenza di un'unica malattia predominante, concentrandosi piuttosto sulla gestione di tutte le condizioni come un insieme [2].

L'aumento dell'aspettativa di vita e l'invecchiamento della popolazione hanno contribuito a un incremento della prevalenza della comorbidità. Questo trend ha portato a un cambiamento nella pratica clinica, con la necessità di passare da un

approccio centrato sulla singola malattia a una gestione olistica delle condizioni del paziente. I medici devono considerare non solo l'interazione tra le malattie stesse, ma anche l'impatto delle terapie su più condizioni contemporaneamente [1].

Oltre agli aspetti clinici, la comorbidità ha conseguenze sociali ed economiche rilevanti. La gestione di pazienti con comorbidità richiede un maggior utilizzo di risorse sanitarie, incluse cure più frequenti e personalizzate, prolungamenti dei tempi di degenza ospedaliera e un impatto diretto sui costi del sistema sanitario [3].

2.2 Social Network Analysis (SNA) nel contesto della comorbidità

La **Social Network Analysis (SNA)** è una tecnica metodologica utilizzata per studiare le relazioni tra entità in una rete. Applicata al campo medico, in particolare nello studio della comorbidità, la SNA permette di visualizzare e analizzare le interazioni tra patologie come se fossero nodi connessi tra loro da archi. Questo approccio aiuta a identificare pattern di comorbidità, ossia insiemi di malattie che tendono a manifestarsi insieme in determinati gruppi di pazienti, e a comprendere come queste patologie si influenzino a vicenda in termini di evoluzione e gravità clinica [4].

In una rete di comorbidità, le malattie più centrali, note come "hub", sono quelle che svolgono un ruolo cruciale nel collegare altre patologie, influenzando così la complessità dei quadri clinici. Le metriche di centralità, come la **betweenness centrality**, che misura quanto un nodo faciliti la comunicazione tra gli altri, e la **degree centrality**, che valuta il numero di connessioni di un nodo, sono essenziali per comprendere quali malattie hanno un impatto maggiore sulla rete di comorbidità. Ad esempio, condizioni come il diabete o le malattie cardiovascolari sono spesso centrali nelle reti di comorbidità, poiché tendono a essere associate a molte altre patologie [5].

L'utilizzo della SNA consente anche di eseguire simulazioni per prevedere l'evoluzione di malattie complesse e supportare la pianificazione delle terapie. Attraverso un'analisi delle connessioni tra le patologie, i ricercatori e i medici possono ottenere una visione più chiara delle interazioni tra malattie e prendere decisioni cliniche più

informate. Ciò rappresenta un grande passo avanti verso una medicina più predittiva e personalizzata [6].

2.3 Visual Analytics nel contesto Clinico

La **Visual Analytics** (VA) rappresenta un'intersezione tra l'analisi dei dati e la visualizzazione interattiva, con l'obiettivo di fornire agli esperti clinici strumenti efficaci per esplorare grandi quantità di dati, come quelli contenuti nei registri elettronici sanitari (EHR). Grazie alla crescente quantità di dati sanitari generati, il ruolo della VA è diventato fondamentale per superare le sfide legate all'**information overload**. Questo sovraccarico di informazioni rende complesso per i medici identificare pattern rilevanti o trarre conclusioni basate su dati non strutturati, specialmente quando si tratta di gestire patologie complesse come le comorbidità.

Nel contesto della **comorbidità**, la VA permette di visualizzare relazioni complesse tra malattie, consentendo ai medici e ai ricercatori di scoprire correlazioni, cluster di patologie e di migliorare la pianificazione terapeutica. Le visualizzazioni interattive forniscono un quadro immediato delle connessioni tra malattie, facilitando il riconoscimento di malattie centrali o "hub" che potrebbero svolgere un ruolo cruciale nel trattamento del paziente [7].

Uno degli aspetti più utili della VA è la sua capacità di combinare analisi statistiche avanzate con rappresentazioni visive intuitive, rendendo più agevole per i medici esplorare i dati e prendere decisioni cliniche più informate. Ad esempio, strumenti di clustering gerarchico permettono di raggruppare i pazienti in base a caratteristiche comuni o a comorbidità, mentre tecniche di filtraggio basate sulla varianza evidenziano le relazioni statisticamente più rilevanti tra patologie. In tal modo, la VA non solo aiuta a esplorare i dati clinici ma supporta anche il processo decisionale attraverso un'interfaccia visiva intuitiva che riduce la complessità delle reti di relazioni [8].

L'integrazione della VA nella ricerca medica e nell'analisi della comorbidità rappresenta una svolta importante per migliorare l'efficienza e la precisione nella diagnosi e trattamento delle malattie croniche complesse.

2.4 Machine Learning e predizione nelle reti di comorbidità

Il **machine learning (ML)** è una branca dell'intelligenza artificiale che consente ai computer di apprendere dai dati senza essere esplicitamente programmati. Attraverso algoritmi complessi, il ML è in grado di individuare pattern nascosti e fare previsioni basate su tali pattern. Le tecniche di ML vengono utilizzate in molte applicazioni, dalla classificazione delle immagini alla predizione di eventi futuri, grazie alla capacità di analizzare grandi quantità di dati e generare modelli predittivi accurati. Tra i metodi più comuni si trovano le reti neurali, gli alberi decisionali, e i modelli di regressione, che sono in grado di adattarsi a problemi molto diversi. Nel contesto delle reti di comorbidità, il machine learning viene applicato per prevedere l'evoluzione delle malattie in pazienti affetti da più condizioni croniche.

Le **reti neurali grafiche (GNN)** rappresentano un'evoluzione recente del machine learning, particolarmente adatta a modellare relazioni complesse come quelle presenti nelle reti sanitarie. Le GNN utilizzano la struttura del grafo, dove i nodi rappresentano pazienti o malattie, e gli archi rappresentano le connessioni tra di essi, come la presenza simultanea di malattie in un paziente o l'associazione tra patologie simili [9].

Applicando le tecniche di ML alle reti di comorbidità, è possibile non solo comprendere le relazioni attuali tra le malattie, ma anche prevedere quali altre condizioni potrebbero svilupparsi in futuro. Studi recenti hanno dimostrato che l'approccio delle GNN può essere utilizzato per effettuare previsioni accurate riguardanti malattie croniche, come il diabete o le malattie cardiovascolari, e per identificare collegamenti latenti tra patologie che potrebbero non essere evidenti con altri metodi [10].

Questi strumenti, insieme ad altre tecniche di machine learning come gli **autoencoder** e l'**embedding delle reti**, stanno migliorando significativamente la capacità predittiva in ambito medico, consentendo di ottimizzare la gestione delle comorbidità e di personalizzare i trattamenti per i pazienti. Il machine learning, quindi, si sta dimostrando un alleato prezioso per la medicina predittiva, fornendo strumenti potenti per anticipare l'insorgenza di nuove malattie e migliorare i risultati clinici [9].

2.5 Graph Neural Networks (GNN) applicate alla predizione delle malattie

Le **Reti Neurali** (neural networks) sono modelli di apprendimento automatico che si ispirano alla struttura del cervello umano, formati da strati di nodi (neuroni) collegati tra loro, in grado di apprendere dai dati e migliorare le loro performance nel tempo. Queste reti vengono utilizzate in diversi ambiti, tra cui la predizione delle malattie, grazie alla loro capacità di modellare e trovare pattern complessi nei dati clinici.

Le **reti neurali grafiche** (Graph Neural Networks - GNN), in particolare, sono un'evoluzione delle reti neurali che operano su dati strutturati sotto forma di grafo. Un grafo è una rappresentazione di dati costituita da nodi e archi, in cui i nodi rappresentano entità (come malattie o pazienti) e gli archi rappresentano le relazioni tra queste entità. Le GNN sono particolarmente utili quando i dati presentano una natura intrinsecamente connessa, come nel caso delle reti di comorbidità, dove diverse malattie possono essere collegate tra loro da relazioni complesse. Le GNN riescono a catturare queste relazioni per migliorare la comprensione e la predizione delle malattie [11].

Le **Heterogeneous Graph Neural Networks (HeteroGNN)** sono progettate per gestire grafi con nodi e relazioni di diverse tipologie, offrendo un potente strumento per modellare contesti complessi come quelli clinici. Grazie alla capacità di rappresentare e analizzare diverse entità cliniche (ad esempio pazienti, malattie, farmaci) e le loro interazioni, le HeteroGNN risultano fondamentali per predizioni cliniche avanzate, migliorando la comprensione delle relazioni tra malattie e supportando la diagnosi e il trattamento delle comorbidità [11].

Gli strati di convoluzione come **SAGEConv** e **GATConv** sono particolarmente rilevanti in questo contesto. Il **SAGEConv** (Sample and Aggregation) sfrutta campionamenti efficienti dai vicini di ciascun nodo per aggregare informazioni da grandi grafi eterogenei, rendendolo adatto per il trattamento di dati clinici su larga scala. Dall'altro lato, il **GATConv** (Graph Attention Convolution) utilizza meccanismi di attenzione per assegnare pesi diversi ai nodi vicini, catturando così meglio le intera-

zioni cruciali tra entità cliniche, come pazienti e trattamenti, basate sull'importanza delle connessioni [12].

Questi metodi consentono di migliorare la capacità delle HeteroGNN nel predire con precisione lo sviluppo di malattie e ottimizzare le strategie terapeutiche, fornendo ai medici strumenti di supporto decisionali sempre più sofisticati, capaci di operare su dati eterogenei e complessi.

2.6 Network Analysis e Machine Learning nel contesto sanitario

L'analisi delle reti e il **machine learning** stanno rivoluzionando il settore sanitario, non solo per la diagnosi, ma anche per la prevenzione delle complicazioni cliniche gravi. Uno degli obiettivi principali di queste tecniche è identificare pattern nascosti nei dati sanitari che possano indicare la progressione verso esiti negativi per il paziente. Questo approccio si è dimostrato particolarmente utile in ambiti come l'oncologia e la cardiologia, dove le malattie tendono a interagire in modi complessi e difficili da rilevare con le tecniche tradizionali.

In oncologia, ad esempio, il machine learning è stato utilizzato per analizzare grandi dataset di pazienti, identificando pattern nelle interazioni tra tumori e altre comorbidità che possono peggiorare l'esito clinico. Le tecniche di **Network Analysis** consentono di visualizzare come diverse malattie possano interagire nel corso del tempo, rivelando legami cruciali tra tumori e patologie croniche. Questo tipo di analisi aiuta a prevenire l'aggravarsi della situazione clinica, suggerendo terapie personalizzate e mirate prima che il paziente sviluppi condizioni critiche [13].

Nel campo della cardiologia, tecniche simili vengono applicate per studiare le relazioni tra malattie cardiovascolari e altre patologie comorbide, come il diabete o l'ipertensione. L'uso combinato di **machine learning** e **SNA** consente di identificare pazienti a rischio, prevedendo episodi cardiaci gravi come infarti o insufficienze cardiache sulla base di dati clinici storici e in tempo reale. Questo approccio proattivo supporta i medici nel monitoraggio continuo dei pazienti, migliorando la gestione delle malattie e prevenendo complicazioni acute [13].

Uno degli sviluppi più interessanti è l'uso di **reti neurali grafiche** (GNN) per predire le interazioni future tra malattie e l'evoluzione della salute del paziente. Le reti GNN possono analizzare l'intera storia clinica di un paziente e prevedere quali condizioni potrebbero svilupparsi sulla base di modelli complessi di comorbidità. Questo è particolarmente utile nella gestione di malattie croniche, dove la previsione accurata delle condizioni future può migliorare la personalizzazione delle cure e ridurre il rischio di peggioramenti improvvisi [13].

2.7 Un caso di studio: comorbidità

Lo studio condotto dai dott. Cavallo, Pagano, De Santis e Capobianco [14] si inserisce nell'ambito dell'analisi della comorbidità attraverso un approccio basato su Big Data e reti di comorbidità. L'uso di dati provenienti dai General Practitioner Records (GPR), che comprendono informazioni sulle prescrizioni mediche e sui dati clinici di routine, consente una visione più ampia delle condizioni cliniche dei pazienti rispetto ai tradizionali Electronic Health Records (EHR), solitamente limitati ai pazienti ospedalizzati per malattie gravi.

Questo aspetto è particolarmente rilevante quando si tratta di condizioni croniche come il diabete, dove la presenza di altre patologie (comorbidità) complica la gestione terapeutica e richiede strategie di prevenzione mirate.

Lo studio ha esaminato un campione di 14.958 pazienti e 1.728.736 prescrizioni raccolte in un arco temporale di 10 anni. I dati riguardavano sia pazienti diabetici che non diabetici, e sono stati utilizzati per creare reti di comorbidità, dove ogni nodo rappresentava un gruppo di diagnosi (codici ICD9-CM) e gli archi indicavano la presenza contemporanea di queste diagnosi nello stesso paziente. Attraverso questo modello di rete, è stato possibile visualizzare e quantificare le associazioni tra diverse malattie e come queste variano in base a fattori come l'età e il sesso del paziente [14].

I risultati dello studio hanno confermato che la comorbidità tende ad aumentare con l'età del paziente e che i pazienti diabetici presentano un pattern di comorbidità significativamente più complesso rispetto ai non diabetici. Questa complessità, rappresentata graficamente, ha permesso di identificare malattie che spesso co-occorrono

nei pazienti diabetici, fornendo così un utile strumento per migliorare la gestione clinica di queste persone [14].

In conclusione, l'approccio di rete utilizzato in questo studio ha dimostrato l'efficacia dei dati di prescrizione come strumento autonomo per l'analisi delle comorbidità, permettendo di anticipare trend epidemiologici e di supportare i processi decisionali dei responsabili delle politiche sanitarie. L'utilizzo di tecniche di **network analysis** applicate ai GPR potrebbe rappresentare una strategia scalabile e generalizzabile, applicabile a popolazioni più ampie, migliorando l'efficienza della sanità pubblica e la prevenzione delle complicanze legate alle malattie croniche [14].

2.8 Big Data

2.8.1 Big Data: una panoramica

Il termine *Big Data* si riferisce a grandi volumi di dati che non possono essere elaborati efficacemente utilizzando metodi tradizionali. I Big Data sono caratterizzati dalle cosiddette "cinque V": Volume, Varietà, Velocità, Veridicità e Valore.

Volume: Si tratta della quantità di dati generati ogni secondo in vari ambiti, dai social media ai dispositivi mobili fino a sensori e sistemi IoT. Il volume è forse l'aspetto più emblematico dei Big Data, poiché si parla di petabyte e zettabyte di informazioni prodotte globalmente ogni anno.

Varietà: I dati provengono da diverse fonti e sono di diversi tipi. Possono essere strutturati (ad esempio, tabelle di database), semi-strutturati (log di server web) o non strutturati (immagini, video, testo libero).

Velocità: La velocità si riferisce alla rapidità con cui i dati vengono generati e devono essere elaborati. Con tecnologie come i sensori e le reti di comunicazione, i dati possono essere raccolti e trasmessi in tempo reale.

Veridicità: Questo concetto si riferisce all'accuratezza e all'affidabilità dei dati raccolti. Non sempre i dati generati sono corretti o completi, quindi la gestione dei Big Data richiede tecniche per garantire la loro qualità.

Valore: Alla fine, la vera sfida dei Big Data è riuscire a estrarre informazioni utili e significative per prendere decisioni strategiche, un compito che richiede strumenti avanzati di analisi e capacità predittive.

L'emergere dei Big Data ha avuto un impatto su numerosi settori, dalla finanza alla logistica, ma uno dei campi in cui il loro utilizzo si è dimostrato più promettente è il settore sanitario.

2.8.2 Big Data nel contesto sanitario

Nel settore sanitario, i Big Data rappresentano una risorsa inestimabile per migliorare la diagnosi, la gestione delle malattie e l'efficienza operativa dei sistemi sanitari. La capacità di analizzare grandi quantità di dati provenienti da cartelle cliniche elettroniche (EHR), sensori medici, test di laboratorio, e fonti esterne come dati socio-economici e comportamentali, consente di creare una visione più completa della salute di un paziente o di una popolazione [15].

In particolare, i Big Data hanno cambiato il modo in cui si affrontano le malattie croniche e complesse, come le comorbidità. Analizzando grandi dataset provenienti da popolazioni diverse, i ricercatori possono identificare pattern di malattia, monitorare l'efficacia dei trattamenti e prevedere esiti clinici futuri. Inoltre, i dati raccolti da dispositivi indossabili e sensori consentono di monitorare i pazienti in tempo reale, migliorando la prevenzione e l'intervento tempestivo.

Uno degli esempi più rilevanti dell'applicazione dei Big Data in sanità è l'uso delle reti comorbide per studiare la co-occorrenza delle malattie, come discusso nel nostro caso di studio sulla comorbidità [14]. Grazie alla combinazione di dati clinici e tecniche di analisi di rete, è possibile identificare gruppi di malattie che tendono a manifestarsi insieme, migliorando la capacità di predire l'evoluzione della salute di un paziente e personalizzare il trattamento.

Oltre alle applicazioni cliniche, i Big Data consentono alle autorità sanitarie di prendere decisioni informate sulla gestione delle risorse, l'allocazione dei fondi e la pianificazione delle politiche sanitarie. Ad esempio, i dati epidemiologici possono essere analizzati per monitorare la diffusione delle malattie infettive e prevedere focolai futuri, consentendo una risposta più efficace e tempestiva.

In sintesi, i Big Data nel contesto sanitario non solo migliorano la qualità delle cure per i singoli pazienti, ma aiutano a ottimizzare l'intero sistema sanitario, rendendolo più efficiente e proattivo.

2.9 Pre-Processing dei Dati

Il *data cleaning* è una fase essenziale in qualsiasi processo di analisi dei dati, soprattutto quando si tratta di dati sanitari. Questa operazione consiste nella rimozione o correzione di dati inconsistenti, duplicati o incompleti, garantendo che i dataset utilizzati per l'analisi siano accurati e affidabili. Nel contesto sanitario, dove le informazioni provengono da molteplici fonti (cartelle cliniche elettroniche, prescrizioni, registri ospedalieri), la pulizia dei dati diventa ancora più critica per prevenire errori nell'analisi e garantire l'integrità delle previsioni basate sui modelli di *machine learning*.

La preparazione dei dataset implica anche la normalizzazione e l'organizzazione dei dati in formati che consentano l'applicazione di tecniche di analisi avanzata, come la *network analysis* e i modelli di *machine learning*. Nel caso della comorbidità, la corretta gestione dei codici diagnostici (come ICD9-CM) e delle prescrizioni mediche è fondamentale per costruire reti di comorbidità affidabili e accurate.

2.10 Data Visualization e manipolazione dei grafi clinici

La *data visualization* è uno strumento cruciale per esplorare e comprendere i pattern nascosti nei dati complessi, specialmente in ambito sanitario. Le tecniche di visualizzazione dei dati consentono di rappresentare graficamente le reti di comorbidità, evidenziando le connessioni tra malattie e pazienti e facilitando l'interpretazione delle interazioni cliniche.

L'uso di grafi, in particolare, è fondamentale per rappresentare visivamente le relazioni tra malattie, rendendo evidenti i nodi centrali (malattie *hub*) e le loro connessioni. Grazie alla visualizzazione grafica, è possibile individuare rapidamente malattie che svolgono un ruolo chiave nella rete di comorbidità e identificare pattern che potrebbero non essere evidenti attraverso l'analisi statistica tradizionale.

Inoltre, la manipolazione dei grafi clinici permette ai ricercatori di esplorare scenari simulativi, modificando le connessioni tra i nodi e osservando come i cambiamenti nelle relazioni tra malattie possano influire sulla salute del paziente. Questi strumenti forniscono un approccio visivo e dinamico all'analisi clinica, migliorando il processo decisionale e facilitando la comunicazione tra medici e ricercatori.

CAPITOLO 3

Medodologia ed Implementazione

3.1 Introduzione

La crescente importanza dei Big Data nel settore sanitario ha trasformato il modo in cui vengono gestiti, analizzati e utilizzati i dati per supportare le decisioni cliniche. Questo fenomeno è particolarmente rilevante nello studio delle comorbidità, che si riferiscono alla coesistenza di più malattie in un singolo paziente. In questo contesto, l'obiettivo del presente capitolo è descrivere il processo metodologico e tecnico che ha portato alla creazione di ComorGraph, una piattaforma progettata per analizzare, visualizzare e prevedere la comorbidità utilizzando tecniche di Social Network Analysis (SNA).

ComorGraph è stata sviluppata non solo come uno strumento di visualizzazione, ma come un sistema avanzato di analisi che consente di esplorare i complessi modelli di comorbidità all'interno di vasti dataset medici. L'intero processo, dalla raccolta dei dati alla loro elaborazione e interpretazione, è stato guidato dalla necessità di fornire un sistema efficiente e scalabile per il settore medico, capace di gestire grandi volumi di dati e supportare i medici nel processo decisionale.

L'implementazione della piattaforma è stata orientata verso scelte tecniche mirate a garantire flessibilità e prestazioni elevate, senza sacrificare la facilità d'uso. In

particolare, il sistema è stato progettato per essere user-friendly e al contempo offrire strumenti analitici potenti, in grado di sfruttare le potenzialità di Neo4j come database a grafo, la reattività di React per l'interfaccia utente, e la robustezza di Python per il backend.

Nel presente capitolo verranno descritti i principali passaggi della progettazione e implementazione di ComorGraph, a partire dai requisiti funzionali fino alla scelta delle tecnologie, con particolare attenzione al processo di ottimizzazione del sistema per gestire le complessità dei dati clinici. Ci sarà infine una sezione dedicata al modulo di Machine Learning che afficherà la piattaforma nello studio delle reti di comorbilità.

3.2 Requisiti Funzionali

Per garantire che la piattaforma ComorGraph soddisfi le esigenze del contesto medico in cui è stata sviluppata, sono stati definiti dei requisiti funzionali specifici. Questi requisiti definiscono cosa il sistema deve fare per supportare l'analisi delle comorbidità in maniera efficiente, permettendo ai medici di visualizzare e analizzare i dati clinici con precisione.

Obiettivi dei Requisiti Funzionali

- **Visualizzazione grafi clinici**: Permettere la rappresentazione visiva delle relazioni tra pazienti, malattie e prescrizioni, con la possibilità di esplorare i dettagli di ogni nodo e arco.
- **Applicazione di metriche SNA**: Fornire metriche di analisi delle reti come: Degree Centrality, Betweenness, Closeness, PageRank, K-Core, essenziali per l'analisi delle comorbidità e per identificare malattie centrali o particolarmente connesse.
- **Interattività e analisi temporale**: Consentire l'analisi dinamica dei dati nel tempo, in modo da osservare come le comorbidità si evolvono per ogni paziente.

- **Gestione e caricamento di database personalizzati**: Dare la possibilità di caricare e cambiare database dinamicamente tramite file CSV strutturati.

Tabella dei Requisiti Funzionali

In seguito riporto una porzione della tabella originale dei requisiti funzionali.

Tabella 3.1: Requisiti Funzionali di ComorGraph

ID	NOME	DESCRIZIONE	PRIORITÀ
RF_01	Visualizzazione Dashboard	L'utente deve poter visualizzare un riepilogo con il numero di pazienti, prescrizioni e malattie.	Alta
RF_02	Visualizzazione Grafo Paziente	L'utente deve poter esplorare il grafo del singolo paziente e la sua storia clinica nel tempo.	Alta
RF_03	Analisi Temporal Slider	L'utente deve poter utilizzare uno slider temporale per analizzare l'andamento clinico del paziente.	Media
RF_04	Visualizzazione Grafo Malattie	L'utente deve poter visualizzare le associazioni tra malattie in un grafo interattivo.	Alta
RF_05	Applicazione Metriche SNA	L'utente deve poter applicare metriche come betweenness, closeness, e k-core sui grafi visualizzati.	Alta

Continua nella pagina seguente.

Tabella 3.2: Requisiti Funzionali di ComorGraph - continuo

ID	NOME	DESCRIZIONE	PRIORITÀ
RF_06	Creazione Dinamica Database	L'utente deve poter caricare CSV formattati per creare un nuovo database di analisi.	Media
RF_07	Switch Database	L'utente deve poter cambiare tra diversi dataset caricati per analisi comparative.	Media
RF_08	Pannello Dettagli Interattivo	L'utente deve poter visualizzare i dettagli dei nodi e degli archi selezionati all'interno dei grafi.	Alta

3.3 Requisiti Non Funzionali

I requisiti non funzionali descrivono le caratteristiche di qualità che la piattaforma ComorGraph deve soddisfare per garantire un'esperienza d'uso ottimale e rispondere alle esigenze tecniche del contesto medico.

Obiettivi dei Requisiti Non Funzionali

- **Facilità di utilizzo**: L'interfaccia utente deve essere intuitiva e semplice da utilizzare anche per medici e ricercatori non tecnici.
- **Efficienza**: Le operazioni di analisi sui grafi devono essere eseguite rapidamente, anche in presenza di grandi volumi di dati.
- **Scalabilità**: Il sistema deve poter gestire un numero crescente di dati senza compromettere le prestazioni.
- **Sicurezza**: Poiché si tratta di dati clinici sensibili, il sistema deve garantire che l'accesso ai dati sia sicuro e che le informazioni vengano crittografate.

Tabella dei Requisiti Non Funzionali

Tabella 3.3: Requisiti Non Funzionali di ComorGraph

ID	NOME	DESCRIZIONE	PRIORITÀ
RNF_01	Facilità di Utilizzo	L'interfaccia deve essere intuitiva e utilizzabile senza una formazione tecnica approfondita.	Alta
RNF_02	Efficienza	Le operazioni di analisi devono essere eseguite rapidamente, anche con dataset di grandi dimensioni.	Alta
RNF_03	Scalabilità	Il sistema deve gestire senza problemi dataset di dimensioni crescenti, fino a milioni di record.	Media
RNF_04	Sicurezza	I dati clinici sensibili devono essere protetti da accessi non autorizzati mediante crittografia.	Alta
RNF_05	Compatibilità	La piattaforma deve essere accessibile tramite i principali browser.	Media

3.4 Motivazione delle Scelte Tecnologiche

Le tecnologie selezionate per lo sviluppo della piattaforma **ComorGraph**: **React**, **Python** e **Neo4j**, rappresentano strumenti moderni e consolidati, ampiamente utilizzati nel settore tecnologico e con una vasta gamma di librerie e strumenti che facilitano lo sviluppo di soluzioni complesse come l'analisi della comorbidità.

React è una libreria frontend moderna e robusta, ampiamente utilizzata per creare interfacce utente dinamiche e interattive. È particolarmente apprezzata per la sua flessibilità, scalabilità e capacità di gestire aggiornamenti efficienti dell'interfaccia, rendendola ideale per la visualizzazione di grafi e dati clinici complessi in modo intuitivo. La sua vasta comunità di sviluppatori e il supporto di nu-

merose librerie ne fanno una scelta naturale per lo sviluppo di applicazioni avanzate.

Python è stato scelto per il backend grazie alla sua semplicità e potenza nel campo della manipolazione dei dati e del machine learning. È dotato di numerose librerie scientifiche (come **Pandas**, **NumPy**, **Scikit-learn**) che permettono di gestire, analizzare e processare grandi dataset sanitari senza compromessi in termini di flessibilità. Python è ampiamente utilizzato nel mondo accademico e industriale, il che lo rende una scelta ideale per integrare tecniche avanzate di analisi dei dati e machine learning.

Neo4j, il database a grafo, rappresenta una scelta eccellente per lo studio della comorbidità, poiché permette di modellare le relazioni complesse tra malattie, pazienti e prescrizioni in modo naturale ed efficiente. Con un ampio set di funzionalità integrate, come il linguaggio di query **Cypher** e la **Graph Data Science Library**, Neo4j fornisce un supporto diretto per l'applicazione di metriche di Social Network Analysis, consentendo analisi rapide e intuitive di reti complesse.

L'integrazione di queste tecnologie, solide e mature, assicura che la piattaforma possa gestire efficientemente l'analisi della comorbidità e offrire un'esperienza utente avanzata e altamente interattiva.

3.5 Architettura della Piattaforma

3.5.1 Database (Neo4j)

Neo4j è uno dei più avanzati database a grafo attualmente disponibili, utilizzato principalmente per modellare dati con relazioni complesse tra entità. A differenza dei tradizionali database relazionali, che rappresentano le relazioni tramite tabelle e join, Neo4j rappresenta i dati in forma di grafo, dove nodi ed archi costituiscono le entità e le loro connessioni dirette. Questa struttura consente una rappresentazione più naturale e intuitiva di dati che presentano molte interconnessioni, rendendolo ideale per l'analisi delle reti e delle strutture complesse, come ad esempio le reti sociali o,

nel nostro caso, le **relazioni tra pazienti, malattie e prescrizioni**.

Potenzialità di Neo4j nella Rappresentazione dei Grafi

1. **Efficienza nel gestire relazioni complesse**: le operazioni che coinvolgono molte relazioni, come il calcolo di collegamenti tra malattie in un contesto di comorbidità, sono molto più rapide rispetto a database relazionali tradizionali.

2. **Visualizzazione immediata delle connessioni**: semplifica notevolmente l'interpretazione delle relazioni tra entità (ad esempio, i collegamenti tra malattie basate su prescrizioni condivise).

3. **Facilità di interrogazione con Cypher**: linguaggio specificamente progettato per interrogare grafi. Cypher consente di eseguire ricerche complesse. La sintassi è molto più intuitiva rispetto a quella SQL, specialmente per domande complesse che richiederebbero numerosi join in un database relazionale.

4. Supporto a **Graph Data Science Library**: Uno dei principali vantaggi di Neo4j è l'integrazione nativa con la Graph Data Science Library (GDSL), un plugin che fornisce una vasta gamma di algoritmi per l'analisi delle reti sociali e la teoria dei grafi. Questo plugin supporta algoritmi di clustering, centralità, analisi delle componenti connesse, rilevamento di comunità, predizione di link, e molto altro. Alcune delle metriche più utilizzate includono: **Betweeness Centrality**, **Closeness Centrality**, **Page Rank**, algoritmo **K-Core**.

Metodologie di Rappresentazione: Dallo Studio Ereditato alla Nuova Strategia

Lo studio iniziale del dott. Cavallo e Giordano[5], che ha ispirato il nostro lavoro, era basato su una rappresentazione della comorbidità, in cui le malattie erano i nodi principali e le prescrizioni condivise rappresentavano gli archi che collegavano queste malattie. In tale struttura, ogni arco conteneva tutte le informazioni relative alle prescrizioni comuni tra le malattie, rappresentando un approccio che, sebbene logico, ha rivelato significativi limiti man mano che la mole di dati aumentava.

Limiti della Strategia Iniziale Con l'aumentare delle dimensioni del dataset, questo tipo di rappresentazione ha cominciato a mostrare diverse inefficienze:

- **Complessità computazionale**: Con il crescere del numero di nodi (malattie) e degli archi (prescrizioni condivise), le query per l'estrazione di informazioni specifiche sono diventate molto onerose dal punto di vista computazionale. L'esecuzione di queste query comportava tempi di attesa prolungati e l'impossibilità di gestire grandi volumi di dati in modo efficiente.
- **Ridotta scalabilità**: Quando il numero di relazioni tra malattie superava determinate soglie, la gestione delle query diveniva impraticabile, rendendo estremamente difficile estrapolare le informazioni utili dai dati. Ogni prescrizione condivisa creava nuovi archi tra le malattie, il che significava che, con un dataset esteso, il numero di relazioni cresceva in maniera esponenziale.

Questa strategia aveva dunque un'efficienza limitata nella gestione di grandi volumi di dati e, ancor più importante, nella possibilità di ricavare informazioni utili e specifiche riguardanti le relazioni tra pazienti, malattie e prescrizioni.

Passaggio alla Nuova Strategia di Rappresentazione Per risolvere le limitazioni sopra citate, è stata adottata una nuova strategia di rappresentazione che ha permesso di separare meglio i dati, facilitando le operazioni di query e visualizzazione. Questa nuova strategia si basa su una struttura che include tre tipi principali di nodi:

1. **Paziente**: Nodo identificato dal campo univoco *codice fiscale assistito*. Questo nodo rappresenta il paziente e permette di mantenere una relazione diretta con le prescrizioni e le malattie.
2. **Malattia**: Nodo identificato dal codice ICD9-CM, rappresenta ciascuna malattia nel dataset. L'informazione relativa alle malattie è stata mantenuta, ma in una struttura che facilita la correlazione con i pazienti e le prescrizioni.
3. **Prescrizione**: Nodo identificato dal *codice prescrizione*, contenente tutte le informazioni necessarie a descrivere i farmaci o le terapie somministrate al paziente.

Questa rappresentazione a tre nodi ha ridotto drasticamente la complessità delle query e ha permesso di mantenere separati i diversi tipi di informazioni in modo che fossero più facilmente accessibili e interpretabili.

Relazioni tra Nodi Le relazioni tra questi nodi sono state gestite attraverso tre tipi principali di archi:

1. **DIAGNOSTICATO_CON**: Una relazione tra un paziente e una malattia, che indica quali malattie sono state diagnosticate al paziente. Questa relazione non si limita a indicare la presenza della malattia, ma include anche informazioni aggiuntive come la *data della prima diagnosi*, *data dell'ultima diagnosi* e il *conteggio delle diagnosi ripetute*.

2. **CURATA_CON**: Questa relazione collega le malattie alle prescrizioni, indicando quali farmaci sono stati utilizzati per trattare specifiche patologie. Questa informazione è fondamentale per analizzare l'efficacia dei trattamenti e per correlare le malattie tra di loro attraverso prescrizioni comuni.

3. **ASSOCIATA_A**: Una relazione tra malattie che rappresenta la comorbidità, ovvero la tendenza di determinate malattie a presentarsi contemporaneamente nello stesso paziente.
   È doveroso citare che la realizzazione di questa relazione è basata sul medesimo studio[5] in cui due malattie sono associate tra di loro quando vengono trattate contemporaneamente attraverso la stessa prescrizione per lo stesso paziente.

Riduzione della Complessità Computazionale La nuova strategia di rappresentazione ha consentito una notevole riduzione della complessità computazionale. Con la separazione dei nodi e la semplificazione delle relazioni, le query per estrarre informazioni specifiche sono diventate molto più rapide ed efficienti. Inoltre, la rappresentazione a tre nodi ha permesso una visualizzazione più intuitiva e comprensibile delle relazioni tra pazienti, malattie e prescrizioni, rendendo più agevole l'analisi dei pattern di comorbidità.

Limiti Hardware e Scelte del Dataset Ridotto Nonostante la nuova struttura di rappresentazione abbia ridotto in modo significativo la complessità delle operazioni sul database, il volume dei dati rimaneva ancora una sfida. Gestire 17 milioni di

record su un database locale ha portato a limitazioni hardware, con difficoltà nella gestione delle risorse e nel tempo di esecuzione delle query. Per questo motivo, è stata presa la decisione di lavorare su un dataset ridotto per la piattaforma, riducendo il numero di record a circa 40.000 entry. Questo dataset è stato creato con un approccio bilanciato che ha mantenuto comunque la rappresentatività dei dati clinici per l'analisi della comorbidità. I dettagli relativi alla creazione di questo dataset verranno trattati nel capitolo seguente.

Vantaggi della Nuova Rappresentazione

Questa rappresentazione ha consentito:

- Una maggiore efficienza nell'elaborazione delle query.
- Un miglioramento delle capacità di visualizzazione e analisi.
- La possibilità di eseguire calcoli di metriche di Social Network Analysis (SNA) direttamente all'interno di Neo4j grazie all'uso della **Graph Data Science Library (GDSL)**, che include metriche essenziali come *degree centrality*, *betweenness*, *closeness*, *PageRank* e *K-core*.

In conclusione, la nuova struttura di rappresentazione ha permesso di ottenere un notevole miglioramento delle prestazioni e dell'efficienza nell'analisi delle relazioni comorbili tra malattie, garantendo al contempo la possibilità di gestire e visualizzare dati clinici complessi in modo ottimizzato.

3.5.2 Backend (Python)

Il backend della piattaforma *ComorGraph* è basato su *Python*, con un'architettura organizzata secondo il pattern *Service-Model-Routes*. Questo approccio permette una chiara separazione delle responsabilità, facilitando la gestione, la manutenzione e l'espansione del codice.

1. Struttura: Service-Model-Routes

Models: Qui vengono definiti i modelli che interagiscono direttamente con il database *Neo4j*. Ogni funzione presente nel modello esegue operazioni *CRUD* (Create, Read, Update, Delete) o query particolari sul database.

- **graph_model.py**: gestisce il recupero dei grafi per pazienti, prescrizioni e malattie. Questo file contiene le query dirette per *Neo4j* che estraggono i dati in base a vari criteri.
- **utils_model.py**: definisce funzioni ausiliarie, come il conteggio di pazienti, prescrizioni e malattie, e metriche di analisi della rete (degree centrality, betweenness, closeness, etc.).

Services: Contiene la logica di business che utilizza i metodi dei modelli per eseguire operazioni specifiche. I servizi agiscono come intermediari tra i modelli e le rotte, incapsulando la logica di alto livello.

- **graph_service.py**: gestisce l'interazione con *graph_model.py*, fornendo servizi per il recupero dei grafi (paziente, prescrizione, malattia).
- **utils_service.py**: si occupa di elaborare i dati di contatori e metriche e fornisce funzionalità di caricamento *CSV* per la creazione dinamica di database.

Routes: Qui si definiscono le *API REST* che gestiscono le richieste *HTTP* da parte del frontend. Le rotte richiamano i servizi e ritornano i dati in formato *JSON*.

- **graph_routes.py**: definisce le rotte per ottenere i grafi per pazienti, prescrizioni e malattie.
- **utils_routes.py**: gestisce le *API* per il caricamento dei *CSV* e per ottenere informazioni generali come i contatori di pazienti, prescrizioni, e malattie.

2. Gestione del Database Neo4j

Il file *extensions.py* si occupa della connessione con *Neo4j*. Qui viene implementata la logica per stabilire la connessione con il database e gestire il passaggio tra database dinamici.

- **init_app**: Inizializza la connessione al database con le credenziali appropriate.
- **switch_database**: Permette di cambiare il database attivo, utile nel contesto della piattaforma per gestire diversi dataset clinici.
- **close**: Chiude la connessione al database.

3. Librerie Utilizzate

Alcune delle librerie fondamentali del backend includono:

Flask: Utilizzato per creare l'*API RESTful* che comunica con il frontend. *Flask* è un micro-framework che consente di gestire facilmente richieste *HTTP*.

Flask-CORS: Gestisce le politiche di *Cross-Origin Resource Sharing*, permettendo al frontend (che potrebbe essere su un dominio differente) di accedere alle risorse del backend senza problemi di sicurezza legati al *CORS*.

Neo4j: Il driver *Python* per connettersi a *Neo4j* e gestire le query a grafo, che vengono eseguite per recuperare dati complessi sulle relazioni di comorbidità.

Werkzeug: Utilizzata per funzioni di sicurezza come *secure_filename*, che garantisce che i file caricati abbiano nomi sicuri.

4. Spiegazione del Flusso di Dati

Il backend interagisce con il database *Neo4j* principalmente attraverso due servizi: *graph_service.py* e *utils_service.py*. Questi servizi chiamano i modelli che eseguono le query *Neo4j*, restituendo al frontend i grafi relativi a pazienti, prescrizioni, malattie e le metriche di analisi (come *betweenness centrality* e *page rank*).

Le rotte definite in *routes* ricevono le richieste *HTTP* dal frontend, richiamano i servizi appropriati, e ritornano i risultati come *JSON*, che saranno utilizzati per la visualizzazione grafica nel frontend.

5. Integrazione con il Frontend

L'integrazione tra frontend (*React*) e backend (*Python*) avviene tramite le *API RESTful*. *React* invia richieste *HTTP GET* o *POST* al backend, che risponde con dati

strutturati. Questi dati vengono poi visualizzati in modo interattivo grazie a librerie come *Vis.js* per la gestione dei grafi, garantendo che il medico possa visualizzare relazioni complesse tra malattie, pazienti e prescrizioni.

Questo approccio, basato su servizi modulari è una chiara separazione tra business logic e gestione dei dati, garantisce una manutenzione facilitata e una scalabilità del sistema, consentendo alla piattaforma di gestire grandi quantità di dati clinici in modo efficiente e strutturato.

3.5.3 Frontend (React)

Struttura e Architettura del Frontend

La struttura del frontend è organizzata in una serie di moduli distinti, che facilitano lo sviluppo collaborativo e la suddivisione delle responsabilità del codice. Le cartelle principali includono:

- **assets**: Questa cartella contiene tutte le risorse statiche utilizzate dall'applicazione, come immagini e icone. La gestione delle risorse in modo centralizzato contribuisce a mantenere l'applicazione organizzata e facilmente scalabile.

- **components**: Il cuore dell'architettura del frontend è rappresentato dai componenti *React*, che comprendono elementi come *GraphComponent* (dedicato alla visualizzazione dei grafi clinici), *Sidebar*, e *DetailsPanel*. Ogni componente segue il principio di responsabilità singola, il che significa che ogni modulo è progettato per eseguire una funzione specifica e ben definita.

- **pages**: Le pagine dell'applicazione includono viste come *Homepage*, *PatientPage*, *GraphPage*, e *PrescriptionPage*. Queste pagine gestiscono le diverse sezioni della piattaforma, connesse tra loro tramite il *React Router*, che permette una navigazione fluida e dinamica tra i vari moduli della piattaforma.

- **services**: I servizi gestiscono le richieste asincrone al backend per il recupero di dati. Questo approccio consente una separazione tra la logica di presentazione e la logica di accesso ai dati, migliorando la manutenibilità del codice.

Gestione della Visualizzazione dei Grafi

Uno dei punti di forza della piattaforma *ComorGraph* è la capacità di visualizzare reti complesse che rappresentano le relazioni tra pazienti, malattie e prescrizioni. Questo è reso possibile grazie a *Vis.js*, libreria leader per la gestione dei grafi.

Il *GraphComponent* gestisce la visualizzazione dei nodi (che rappresentano pazienti, malattie o prescrizioni) e degli archi (che rappresentano le relazioni tra di loro). La logica sottostante è contenuta in *GraphComponentLogic.js*, dove i dati ricevuti dal backend vengono trasformati in strutture grafiche interattive.

Ogni nodo è rappresentato visivamente con un'icona e un colore distintivo per facilitarne l'identificazione immediata. I nodi paziente, malattia e prescrizione vengono rappresentati con colori distinti: rosso (Paziente), blu (Malattia) e verde (Prescrizione). Questa scelta facilita la comprensione delle relazioni tra i dati clinici in modo visivo e intuitivo.

Il grafo è dinamico e interattivo: l'utente può zoomare, spostarsi e cliccare sui nodi e archi per ottenere dettagli aggiuntivi. Inoltre, è disponibile la funzionalità per calcolare le metriche *SNA* (*Social Network Analysis*) e applicarle sul grafo corrente, che viene aggiornato ridimensionando i nodi malattia in base al valore della metrica desiderata.

Gestione delle Richieste al Backend

La comunicazione tra il frontend e il backend è gestita da una serie di *service* definiti in file separati, come *graphDataService.js* e *utilsDataService.js*. Questi servizi utilizzano il metodo *fetch* per inviare richieste al backend e recuperare i dati necessari per la visualizzazione e l'interazione nell'interfaccia utente.

- **graphDataService**: gestisce le richieste per ottenere il grafo completo o i singoli grafi per pazienti, malattie e prescrizioni.
- **utilsDataService**: si occupa di fornire i risultati delle metriche di *SNA* e della gestione dei dati generali, come: il caricamento di nuovi file *CSV*, la creazione dinamica del database e il recupero delle statistiche globali dell'applicazione.

Vite come Strumento di Build

Il frontend di *ComorGraph* è stato sviluppato utilizzando *Vite*, un moderno strumento di build per applicazioni web. *Vite* è stato scelto in quanto offre numerosi vantaggi rispetto a strumenti più tradizionali come *Create React App* (CRA):

Velocità di sviluppo: *Vite* utilizza *ESBuild* per gestire la fase di sviluppo, il che si traduce in tempi di build e reload significativamente più rapidi, specialmente per progetti di grandi dimensioni. Permette di effettuare modifiche real-time, salvare ed effettuare la compilazione solo del modulo richiesto, visualizzando quasi immediatamente i risultati della modifica.

Ottimizzazione del codice: Grazie al suo sistema di bundling, *Vite* ottimizza il codice per la produzione in modo più efficiente, riducendo i tempi di caricamento e migliorando le performance dell'applicazione.

Supporto avanzato per le librerie: *Vite* supporta nativamente molte librerie moderne e permette una configurazione flessibile tramite il file *vite.config.js*, consentendo una personalizzazione ottimale del processo di build.

In questo contesto, *Vite* è stato scelto non solo per la sua rapidità ma anche per la sua capacità di gestire applicazioni con un elevato numero di dipendenze, come quella di *ComorGraph*, che si affida a numerose librerie per la gestione grafica e la visualizzazione dei dati clinici.

Con questo approccio, la piattaforma *ComorGraph* offre una soluzione robusta e scalabile per l'analisi della comorbidità, utilizzando le migliori tecnologie per garantire performance elevate, una user experience fluida, e l'interattività richiesta per l'analisi di dati clinici complessi.

3.6 Modulo di IA (HeteroGNN)

L'analisi delle reti complesse, soprattutto in ambito sanitario, ha visto un incremento significativo nell'adozione di tecniche avanzate per studiare la comorbidità, ossia l'interazione tra malattie nei pazienti affetti da patologie multiple. Le **Reti Neurali Grafiche** (Graph Neural Networks, GNN) rappresentano oggi una delle tecniche di punta per modellare tali interazioni, poiché permettono di studiare i dati strutturati sotto forma di grafi, dove le entità cliniche (ad esempio, pazienti e malattie) sono rappresentate come nodi e le relazioni tra di esse come archi.

Tuttavia, un ulteriore passo avanti è stato fatto con l'introduzione delle **Heterogeneous Graph Neural Networks (HeteroGNN)**, che consentono di modellare reti eterogenee composte da diversi tipi di nodi e relazioni. Questo approccio risulta particolarmente utile in contesti complessi come quello della sanità, dove diverse entità (pazienti, malattie, prescrizioni) interagiscono tra loro in modo intricato.

Nell'ambito di questo progetto, pur non approfondendo tecnicamente i dettagli implementativi delle HeteroGNN, si rimanda alla tesi del collega Tullio Mansi per uno studio dettagliato su come le HeteroGNN possano essere utilizzate per predire le relazioni paziente-malattia, basandosi su dati clinici e cronologie di diagnosi.

3.6.1 Modelli di HeteroGNN

Due modelli principali di HeteroGNN sono stati implementati e confrontati per analizzare la comorbidità:

1. **Modello con SAGEConv**: Basato su convoluzioni eterogenee, questo modello utilizza un'aggregazione dei nodi vicini per costruire rappresentazioni grafiche delle entità. Questo approccio si è rivelato utile per individuare correlazioni dirette tra pazienti e malattie, offrendo una rappresentazione basilare ma efficace del grafo.

2. **Modello con GATConv**: Questo modello introduce un meccanismo di attenzione (Graph Attention Networks) per assegnare pesi differenti ai nodi vicini in base alla loro rilevanza. Il modello aggiunge anche una componente temporale, introducendo archi che rappresentano la progressione cronologica delle

diagnosi tra pazienti. Questo rende la rappresentazione più dettagliata e adatta a cogliere l'evoluzione clinica delle malattie nel tempo.

Per i dettagli implementativi e una valutazione approfondita di questi modelli, si rimanda nuovamente alla tesi del collega, che esamina le diverse architetture e le loro performance attraverso metriche standard come **accuracy**, **ROC AUC**, **precision**, **recall** e **f1-score**.

3.6.2 Predizione di Relazioni Paziente-Malattia

Il compito principale delle **HeteroGNN** applicate a questo contesto è la predizione delle probabilità che un paziente sviluppi una determinata malattia, basandosi sulle informazioni di diagnosi pregresse e sulla struttura della rete. Attraverso l'uso di tecniche di **train-test split** e **negative sampling**, si è potuto bilanciare il dataset, migliorando la capacità dei modelli di fare previsioni accurate.

Il confronto tra i modelli basati su **SAGEConv** e **GATConv** ha mostrato che l'inclusione di relazioni temporali tra pazienti migliora la capacità predittiva del modello, permettendo di catturare l'evoluzione clinica in maniera più accurata.

Conclusione

È importante sottolineare come tali reti rappresentino una componente cruciale nella comprensione e predizione delle comorbidità. Le tecniche di **Social Network Analysis** forniscono una base teorica che, combinata con l'analisi delle **HeteroGNN**, può aprire nuove prospettive nell'applicazione dell'intelligenza artificiale in ambito sanitario.

L'analisi congiunta dei risultati ottenuti tramite l'uso delle GNN per la predizione delle relazioni paziente-malattia permette di arricchire ulteriormente il panorama delle tecniche di analisi e predizione nell'ambito delle comorbidità, migliorando la comprensione delle dinamiche delle patologie e offrendo nuovi spunti per la medicina.

CAPITOLO 4

Pre-Processing e Bilanciamento del Dataset Clinico

4.1 Introduzione

Obiettivo dell'analisi e del dataset bilanciato

L'obiettivo primario di questa analisi è creare un dataset bilanciato che consenta un'analisi accurata delle relazioni di comorbilità tra diverse malattie, sfruttando metriche di *Social Network Analysis* (SNA) e modelli di *machine learning*. Un dataset bilanciato è essenziale per evitare distorsioni derivanti dalla sovra-rappresentazione o sotto-rappresentazione di alcune categorie di dati, che potrebbero compromettere l'affidabilità dei risultati. Garantendo un bilanciamento ottimale dei dati, si ottiene una base solida per lo sviluppo di modelli predittivi che possano fare previsioni generalizzabili, senza essere influenzati da squilibri strutturali.

La robustezza dei modelli dipende in gran parte dalla qualità e dalla rappresentatività del dataset. Un dataset bilanciato riduce il rischio di *bias* sistematici, migliorando così l'affidabilità delle analisi cliniche e permettendo una rappresentazione fedele dei fenomeni di comorbilità. Questo bilanciamento, inoltre, facilita l'efficienza computazionale, elemento cruciale in un contesto di risorse hardware limitate. Infatti, un ulteriore scopo del bilanciamento è quello di ottimizzare le risorse computazionali

disponibili, dato che la piattaforma, in questa fase, opera su hardware limitato.

Creare un dataset pulito e bilanciato consente di ridurre il carico computazionale, migliorare i tempi di esecuzione delle query e rendere il sistema scalabile. In una fase successiva, il dataset sarà trasferito su un'infrastruttura più potente, permettendo di sfruttare capacità di calcolo maggiori e analizzare dataset di dimensioni ancora più ampie.

4.2 Fase 1: Data Collection

I dati utilizzati per l'analisi derivano dalla collaborazione con cinque diverse ASL della regione Campania, sotto la supervisione del dott. Pierpaolo Cavallo. Questa raccolta di informazioni cliniche è stata fondamentale per costruire un dataset solido che riflettesse le comorbilità tra malattie e le relative prescrizioni. I dati coprono diverse aree, tra cui prescrizioni mediche, dati di fragilità, dati anagrafici e informazioni sui medici, ciascuna delle quali è stata suddivisa in file CSV specifici. È importante osservare già da adesso che tutte le informazioni sensibili dei pazienti e medici sono già state fornite in forma **anonima** dalla sorgente, motivo per cui non è stata necessaria alcuna crittografia dei dati ricevuti.

Il dataset è stato raccolto in un ampio intervallo temporale, con Medservice 5 che si distingue per la segmentazione in fasce di età decennali e una copertura temporale dal 1900 al 2019. Tuttavia, i dati di fragilità per questo dataset si fermano al 1959, una limitazione considerata nella successiva fase di data cleaning. Grazie alla varietà e alla portata dei dati, il dataset fornisce una base adeguata per lo studio delle comorbilità e consente di indagare su molteplici fattori che possono influenzare la salute dei pazienti.

È opportuno puntalizzare che partire dal **2004**, le ASL della Campania hanno iniziato a digitalizzare le prescrizioni mediche come parte dell'implementazione del **Decreto Legge n. 269/2003**, convertito nella **Legge n. 326/2003**, che ha istituito il **Sistema Tessera Sanitaria** e promosso la digitalizzazione delle prescrizioni mediche. Questo decreto ha incoraggiato le strutture sanitarie a transitare ai sistemi informativi, un processo che è stato poi reso obbligatorio con il **Decreto Legge n. 179/2012**.

Quest'ultimo ha accelerato l'introduzione delle ricette elettroniche e la progressiva sostituzione delle prescrizioni cartacee con quelle digitali.

I dati antecedenti al 2004 sono dunque soggetti a un numero significativo di errori, in quanto venivano spesso compilati manualmente e successivamente trasferiti in formato digitale, con la tendenza a riempire campi nulli in modo poco coerente e con tecniche automatiche. Questa gestione inefficiente dei dati storici rende complessa l'analisi, ma rappresenta anche una sfida nel tentativo di estrarre il massimo valore da tali dati attraverso un accurato processo di data cleaning.

Struttura del dataset

Il dataset è suddiviso in quattro categorie principali: prescrizioni mediche, dati anagrafici, dati di fragilità e dati dei medici. Ogni categoria è caratterizzata da campi specifici necessari per garantire un'analisi approfondita delle relazioni cliniche.

Prescrizioni mediche Questa categoria contiene informazioni sui trattamenti prescritti ai pazienti. I campi principali includono:

Tabella 4.1: Struttura Dataset Prescrizioni

NOME	DESCRIZIONE
Codice Regionale Medico	Codice identificativo del medico che ha emesso la prescrizione.
Codice Fiscale Assistito	Codice anonimo del paziente per identificare il paziente garantendo privacy.
Data Prescrizione	Data della prescrizione medica.
Luogo Prescrizione	Luogo in cui è stata effettuata la prescrizione.
Fascia Prescrizione	Fascia di appartenenza della prescrizione.

Continua nella pagina seguente.

Tabella 4.2: Struttura Dataset Prescrizioni - continuo

NOME	DESCRIZIONE
Tipo Prescrizione	Tipo specifico della prescrizione (es. farmaco, dispositivo medico, terapia).
ICD9-CM	Codice diagnostico per la classificazione della malattia trattata.
Data Prima Diagnosi	Data della Prima diagnosi di una malattia di un paziente.
Codice Prescrizione	Identificativo della prescrizione.
AIC	Codice identificativo del farmaco prescritto.
Descrizione Prescrizione	Descrizione dettagliata del farmaco o trattamento.
Quantità Prescrizione	Quantità del farmaco o trattamento prescritto.
Note AIFA	Informazioni supplementari fornite dall'AIFA (Agenzia Italiana del Farmaco).
Codice Esenzione	Codice di esenzione applicabile al paziente.
Prezzo	Costo del farmaco o del trattamento.
Suggerita	Indicazione se la prescrizione è stata suggerita o meno.
Durata Prescrizione	Durata del trattamento prescritto.
Esito	Risultato o esito della terapia prescritta.

Dati di fragilità Questi dati forniscono informazioni dettagliate sulle condizioni fisiche e sociali dei pazienti, con i seguenti campi principali:

Tabella 4.3: Struttura Dataset Fragilità Paziente

NOME	DESCRIZIONE
Codice Fiscale	Identificativo in forma anonima del paziente, collegabile ai dati anagrafici e alle prescrizioni.
Anno Nascita	Anno di nascita del paziente, utile per l'analisi demografica.
Sesso	Genere del paziente.
Patologia	Patologia principale che caratterizza la fragilità del paziente.
Terapia	Terapia in corso per gestire la condizione di fragilità.
Condizione Sociale	Condizione sociale del paziente, utile per valutare il supporto esterno.
Score Sensorio Comunicazione	Valutazione delle capacità sensoriali e comunicative del paziente.
Score Cognitivo	Misura delle capacità cognitive del paziente.
Score Mobilita	Valutazione del livello di mobilità del paziente.
Score Situazione Funzionale	Indicatore dello stato funzionale complessivo del paziente.

Dati anagrafici Questa categoria descrive le caratteristiche demografiche e geografiche dei pazienti, inclusi i seguenti campi:

Tabella 4.4: Struttura Dataset Anagrafica Pazienti

NOME	DESCRIZIONE
Codice Fiscale	Identificativo in forma anonima del paziente per correlare le informazioni cliniche e anagrafiche.
Anno Nascita	Anno di nascita del paziente.
CAP	Codice di avviamento postale che identifica la residenza del paziente.
Sesso	Genere del paziente.
Latitudine e Longitudine	Coordinate geografiche della residenza del paziente. Saranno eliminate per preservare la privacy.

Dati dei medici: Include informazioni sui medici che hanno effettuato le prescrizioni:

Tabella 4.5: Struttura Dataset Medici

NOME	DESCRIZIONE
Codice Regionale	Identificativo del medico all'interno del sistema sanitario regionale.
Anno Nascita	Anno di nascita del medico.
Sesso	Genere del medico.

4.3 Fase 2: Data Cleaning

Il data cleaning rappresenta una fase cruciale per garantire l'affidabilità del dataset, eliminando dati incompleti, errati o non rilevanti. Questo processo ha permesso di costruire un dataset più pulito e coerente, assicurando che le analisi successive fossero basate su dati validi e consistenti. Durante questa fase, sono state applicate diverse operazioni, tra cui l'eliminazione dei campi nulli, la correzione di errori nei codici diagnostici e la rimozione di righe duplicate. Ecco i passaggi chiave del processo:

Eliminazione dei campi nulli

La presenza di campi nulli, specialmente in variabili fondamentali come il codice diagnostico ICD9-CM, il codice fiscale anonimizzato del paziente o il codice prescrizione, avrebbe compromesso l'affidabilità del dataset. Per evitare distorsioni o perdite di informazioni rilevanti, è stato essenziale eliminare tutte le righe contenenti valori mancanti in questi campi chiave.

L'operazione è stata eseguita utilizzando la funzione `dropna()` della libreria Pandas, che ha permesso di individuare e rimuovere in modo efficiente le righe con dati mancanti nei campi critici. Questo passaggio ha garantito che nel dataset rimanessero solo record completi e utilizzabili per l'analisi delle comorbilità, migliorando l'affidabilità delle successive analisi cliniche.

Correzione degli errori nei codici ICD9-CM

Durante la fase di pulizia dei dati, è emerso un problema ricorrente nella colonna ICD9-CM, dove la lettera "O" era stata erroneamente inserita al posto del numero "0". Questo errore avrebbe potuto portare a un'errata classificazione delle malattie, compromettendo l'accuratezza delle analisi. Dopo una verifica con fonti scientifiche e cliniche autorevoli, si è confermato che quei codici contenenti "O" non esistevano e che si trattava di errori di digitazione.

L'operazione di correzione è stata eseguita utilizzando la funzione `replace()` di Pandas, che ha permesso di sostituire ogni occorrenza della lettera "O" con lo zero "0", garantendo la correttezza dei codici diagnostici utilizzati nel dataset. Questa

correzione ha migliorato la precisione complessiva del dataset, eliminando codici non validi e rafforzando la qualità dei dati per l'analisi delle comorbilità.

Correzione degli errori di Genere

Durante questa fase così come per i codici ICD9-CM, sono stati rilevati alcuni errori di digitazione nella colonna "Sesso" che sono stati risolti eliminando l'intera riga perchè consistevano in poche occorrenze che non avrebbero condizionato lo studio. Questi errori sono visualizzabili nei grafici del sottoparagrafo 4.8.1.

Rimozione delle righe duplicate

Un ulteriore passaggio chiave del data cleaning ha riguardato la rimozione delle righe duplicate. I duplicati possono generare distorsioni nelle analisi statistiche e nei modelli predittivi, aumentando il peso di determinati record e portando a stime gonfiate o imprecise. Per evitare questa problematica, tutte le righe duplicate sono state identificate e rimosse.

La funzione `drop_duplicates()` di Pandas è stata utilizzata per eseguire questa operazione, identificando righe duplicate basate su campi chiave come il codice fiscale anonimizzato, il codice ICD9-CM e la data di prescrizione. Questo passaggio ha garantito che ogni record rappresentasse un evento clinico unico e distintivo, senza influenze dovute a ridondanze o ripetizioni non intenzionali.

4.4 Fase 3: Data Transformation

Dopo la fase di **Data Cleaning**, in cui il dataset viene pulito da errori, valori mancanti e duplicati, il passo successivo consiste nella **Data Transformation**. Questa fase è fondamentale per ottimizzare il dataset in vista delle analisi e dei modelli predittivi. L'obiettivo principale della **data transformation** è *uniformare* e *standardizzare* le variabili per garantire coerenza e comparabilità tra i dati. Questo processo riduce la variabilità non controllata, specialmente quando si lavora con variabili che hanno scale diverse o rappresentano tipi di dati differenti, come date o numeri.

La **data transformation** comprende tecniche come la *normalizzazione*, il *filtraggio* e il *ridimensionamento* delle feature, operazioni che assicurano una rappresentazione omogenea delle informazioni. Questi passaggi sono particolarmente importanti nei dataset clinici, dove la coerenza tra variabili temporali e numeriche gioca un ruolo cruciale per ottenere risultati affidabili.

Normalizzazione delle date

La normalizzazione delle date è stata una delle operazioni più importanti per uniformare le variabili temporali all'interno del dataset. In un dataset clinico, le date di prescrizione e le date di diagnosi sono fondamentali per tracciare l'evoluzione delle malattie e l'interazione tra trattamenti medici. Tuttavia, le date possono essere registrate in formati differenti, creando ambiguità e difficoltà nelle analisi temporali.

Per risolvere questa problematica, tutte le date presenti nel dataset sono state normalizzate utilizzando il formato ISO 8601. Questo standard internazionale garantisce una rappresentazione coerente delle informazioni temporali e permette un confronto agevole tra date provenienti da differenti fonti o sistemi. Il formato ISO 8601 (`YYYY-MM-DD`) è stato scelto per la sua adozione universale e per la capacità di ridurre ambiguità, soprattutto in contesti internazionali o clinici, dove le convenzioni di data possono variare.

Filtraggio temporale (1960-2010)

Una delle decisioni chiave durante il processo di normalizzazione è stata l'applicazione di un filtro temporale che limitasse i dati alle prescrizioni e alle diagnosi comprese tra il 1960 e il 2010. Questo intervallo temporale è stato scelto per garantire la coerenza dei dati analizzati, escludendo periodi con pratiche mediche e condizioni socio-economiche troppo differenti, che avrebbero potuto introdurre variabilità indesiderata nell'analisi delle comorbilità.

Il periodo precedente al 1960 è stato escluso poiché i trattamenti medici, le diagnosi e la disponibilità di farmaci differivano notevolmente rispetto alle pratiche attuali. L'inclusione di dati antecedenti avrebbe potuto influenzare negativamente la qualità delle analisi, creando disomogeneità tra i periodi storici. Allo stesso modo, si è

deciso di escludere anche i dati successivi al 2010 per mantenere una certa uniformità nei trattamenti analizzati, evitando di includere dati troppo recenti che riflettono pratiche mediche non ancora consolidate.

L'uso di questo intervallo temporale ha permesso di mantenere solo i dati clinici più rilevanti e coerenti, migliorando la capacità di effettuare analisi accurate e prive di distorsioni storiche.

Filtro sul Numero di Prescrizioni per Paziente

Un ulteriore aspetto fondamentale è stato il bilanciamento basato sul numero di prescrizioni per paziente. Pazienti con un numero estremamente basso o estremamente alto di prescrizioni potrebbero distorcere l'analisi, poiché potrebbero rappresentare casi atipici o anomali.

Per questo motivo, sono stati esclusi dal dataset:

- **Pazienti con meno di 11 prescrizioni**: I pazienti con un numero troppo basso di prescrizioni potrebbero rappresentare casi clinici non completi o non significativi, che non fornirebbero un contributo sufficiente per identificare le comorbilità in modo accurato.

  Il valore di 11 è stato selezionato in seguito a un'analisi della distribuzione dei dati, con l'obiettivo di mantenere solo quei pazienti con un'interazione medica sufficiente da permettere un'analisi delle comorbilità clinicamente rilevante. Pazienti con un numero limitato di prescrizioni avrebbero fornito un numero di dati insufficiente, generando rumore nell'analisi e riducendo la qualità dei risultati.

- **Pazienti con più di 1000 prescrizioni**: Questi pazienti sono stati considerati *outliers*, ovvero casi estremi che avrebbero potuto influenzare negativamente le analisi statistiche e i modelli predittivi. Un numero così elevato di prescrizioni per paziente non riflette la tipica interazione medica della popolazione e rischia di introdurre distorsioni nei risultati.

Filtro sulle occorrenze di Malattie

Per garantire che l'analisi delle comorbilità si concentrasse su malattie clinicamente rilevanti, sono state escluse tutte le malattie con meno di **4 occorrenze** nel dataset. Le malattie troppo rare, infatti, avrebbero introdotto variabilità eccessiva, senza fornire un contributo significativo all'analisi globale. Inoltre, la presenza di patologie rare avrebbe potuto aumentare il rischio di **overfitting** nei modelli di machine learning, riducendo la loro capacità di generalizzazione.

La soglia di 4 occorrenze è stata scelta sulla base di un'analisi statistica preliminare che ha identificato un equilibrio tra rappresentatività clinica e stabilità dei dati. Questa selezione è supportata dalla letteratura, che suggerisce l'eliminazione delle classi con bassa frequenza per ridurre la variabilità non necessaria e migliorare l'efficacia dei modelli predittivi [16].

Feature Scaling

È stato importante applicare tecniche di feature scaling per omogeneizzare i dati numerici. Alcune variabili, come la quantità di prescrizione, si estendono su scale diverse rispetto ad altre informazioni contenute nel dataset. L'uso di feature scaling ha garantito che tutte le variabili fossero proporzionate in modo adeguato, migliorando così l'accuratezza dei modelli predittivi e delle analisi statistiche.

Il ridimensionamento è stato eseguito utilizzando tecniche come la *min-max normalization*, che riduce i valori numerici a un intervallo predeterminato, migliorando la capacità dei modelli di machine learning di trattare i dati in modo efficace.

4.5 Fase 4: Data Integration

Dopo la fase di data cleaning, è stato necessario arricchire il dataset delle prescrizioni con informazioni demografiche provenienti da un dataset anagrafico. Il processo di **Data Integration** è stato eseguito utilizzando il codice fiscale anonimizzato del paziente come chiave di collegamento tra i dataset.

L'integrazione ha permesso di aggiungere nuove variabili demografiche che forniscono un livello di dettaglio importante per l'analisi delle comorbilità.

4.6 Fase 5: Feature Selection

La fase di Feature Selection rappresenta un momento cruciale nel processo di preparazione del dataset, poiché determina quali variabili saranno mantenute per l'analisi delle comorbilità. La selezione accurata delle feature è essenziale per garantire che il dataset finale sia bilanciato, coerente e rappresenti al meglio le informazioni necessarie per il contesto specifico dell'analisi.

Questa fase ha comportato la rimozione di alcune feature iniziali che, dopo un'attenta valutazione, non risultavano rilevanti per lo studio della comorbilità. Sono state scelte solo le feature fondamentali per l'analisi clinica e la costruzione dei modelli predittivi, eliminando campi non utili o ridondanti.

Descrizione della Feature Selection

Inizialmente, il dataset contenente le prescrizioni includeva una vasta gamma di feature, molte delle quali non rilevanti per l'analisi della comorbilità. Il processo di feature selection è stato guidato dall'obiettivo di semplificare il dataset, mantenendo solo le variabili indispensabili per comprendere le relazioni tra pazienti, diagnosi e prescrizioni. Questo processo ha contribuito a ridurre la complessità e migliorare la qualità del dataset.

Feature Prescrizioni Mantenute

Dopo il data cleaning, sono state selezionate solo le seguenti feature dal dataset delle prescrizioni, che sono state giudicate essenziali per lo studio della comorbilità:

Tabella 4.6: Feature Prescrizione Mantenute e Motivazione

NOME	MOTIVAZIONE
Codice Regionale Medico	Utile per identificare i medici che hanno curato i pazienti e le diagnosi associate.
Codice fiscale Assistito	Essenziale per collegare le prescrizioni e le diagnosi ai pazienti mantenendo la privacy.

Tabella 4.7: Feature Prescrizione Mantenute e Motivazione - continuo

NOME	MOTIVAZIONE
Data Prescrizione	Importante per tracciare l'evoluzione temporale delle diagnosi e delle terapie.
Tipo Prescrizione	Utile per analizzare la natura delle cure somministrate.
ICD9-CM	Cruciale poiché permette di collegare patologie e studiare le loro correlazioni.
Data Prima Diagnosi	Campo rilevante per comprendere la progressione delle malattie nel tempo.
Codice Prescrizione	Serve a identificare in modo specifico ciascuna prescrizione.
Descrizione Prescrizione	Utile per analisi più dettagliate sulle terapie associate alle diagnosi.
Quantità Prescrizione	Importante per valutare l'intensità e la durata delle cure.

Queste feature sono state considerate fondamentali per il contesto clinico, in quanto rappresentano le informazioni chiave necessarie per comprendere le interazioni tra pazienti, malattie e prescrizioni mediche.

Feature Prescrizioni Eliminate

Le feaure non selezionate per l'analisi includevano variabili che, pur presenti nel dataset originale, non erano direttamente pertinenti per lo studio della comorbilità o erano incomplete. Alcune di queste variabili includevano un numero elevato di dati nulli o non fornivano informazioni rilevanti per l'analisi. Le feature eliminate includono:

Tabella 4.8: Feature Prescrizione Rimosse e Motivazione

NOME	MOTIVAZIONE
Luogo Prescrizione	Non direttamente rilevante per lo studio effettuato.
Fascia Prescrizione	Informazione di dettaglio quasi sempre auto-compilata con valore 1.
AIC	Codice identificativo del farmaco, era spesso mancante e difficile trovare pattern affidabili per calcolarlo.
Note AIFA	Informazione non rilevante per il contesto dell'analisi.
Codice Esenzione	Informazione non rilevante per il contesto dell'analisi, spesso mancante.
Prezzo	Informazione non rilevante per il contesto dell'analisi.
Suggerita	Informazione non rilevante per il contesto dell'analisi.
Durata Prescrizione	Informazione parsa come auto-compilata, quasi sempre pari ad 1.
Esito	Informazione mancante per quasi l'interezza del dataset.

Feature Anagrafiche Mantenute

Delle feature ottenute con il dataset anagrafico, sono state mantenute solo:

Tabella 4.9: Feature Anagrafiche Mantenute e Motivazione

NOME	MOTIVAZIONE
Sesso	Importante per differenziare l'analisi tra pazienti maschi e femmine e osservare eventuali differenze di comorbilità tra i sessi.
CAP	Permette di analizzare eventuali correlazioni tra luogo di residenza e comorbilità.
Anno Nascita	Fattore critico per analizzare le comorbilità, poiché molte malattie si manifestano in diverse fasce di età.

Queste feature sono risultate fondamentali per aggiungere un livello di dettaglio demografico all'analisi delle comorbilità, permettendo di considerare anche fattori come l'età e il sesso, che influenzano l'incidenza delle malattie. L'inserimento delle feature riguardanti la residenza erano informazioni che sono state rimosse per preservare la privacy dei pazienti e per evitare che il modello di machine learning potesse trovare dei pattern troppo specifici e andare in overfit.

Eliminazione Feature Fragilità

In questa parte di studio si è deciso, vista la complessità dei dati relativi alle fragilità e le numerose incoerenze, con score difficili da interpretare e spesso dati mancanti, di non considerare queste feature e rimandare a studi futuri l'integrazione.

Nuove Feature Aggiunte

Nel contesto della piattaforma *ComorGraph*, sono state aggiunte due nuove feature essenziali per migliorare l'efficienza computazionale e fornire informazioni più immediate all'utente:

Tabella 4.10: Feature Aggiunte e Motivazione

NOME	MOTIVAZIONE
Data Ultima Diagnosi	Rappresenta la data più recente in cui è stata identificata una prescrizione per uno specifico codice malattia in un paziente. Questo campo è stato calcolato automaticamente per ogni paziente e per ogni malattia, consentendo una rapida consultazione del dato senza dover eseguire costosi calcoli computazionali in tempo reale.
Descrizione Malattia	Fornisce una descrizione testuale del gruppo di malattie a cui appartiene il codice ICD9-CM. Questa feature è stata introdotta per semplificare l'interpretazione dei dati, soprattutto per coloro che non hanno familiarità con i codici ICD9-CM, permettendo un'analisi rapida delle comorbidità senza dover fare riferimento a tabelle di corrispondenza.

4.7 Fase 6: Data Balancing

Il bilanciamento dei dati è una fase cruciale per garantire che il dataset rappresenti in modo equo le diverse fasce di popolazione. Nel contesto dell'analisi delle comorbilità e dell'addestramento di modelli di machine learning, un dataset non bilanciato potrebbe introdurre bias, portando a risultati distorti e non rappresentativi. Il bilanciamento, pertanto, è fondamentale per correggere eventuali squilibri nelle variabili demografiche e cliniche, migliorando la qualità e l'accuratezza delle analisi e dei modelli predittivi.

Resampling Per bilanciare il dataset, è stato utilizzato il **resampling**, una tecnica che modifica la distribuzione di una classe aumentando o riducendo il numero di campioni presenti. L'obiettivo del resampling è ottenere una distribuzione equilibrata, correggendo eventuali disuguaglianze tra classi.

In questo studio, il resampling è stato applicato per correggere gli squilibri in due dimensioni principali: **fasce d'età**, e **sesso**. Due tecniche di resampling sono state impiegate a seconda del contesto:

- **Oversampling**: Aumento dei campioni della classe minoritaria, duplicando i dati esistenti o generando nuovi campioni sintetici. Questa tecnica è stata utilizzata quando una classe risultava sottorappresentata, come nel caso di fasce d'età o gruppi di sesso meno numerosi.
- **Undersampling**: Riduzione dei campioni della classe maggioritaria, eliminando parte dei dati per riequilibrare la distribuzione. Questa tecnica è stata applicata quando una classe risultava sovra-rappresentata, ad esempio nel caso di pazienti con un numero eccessivo di prescrizioni rispetto alla media.

L'applicazione del resampling ha permesso di ottenere un dataset che riflette in modo più equo le diverse caratteristiche della popolazione, riducendo il rischio di distorsioni nelle analisi e migliorando la robustezza dei modelli predittivi.

Bilanciamento per Fasce d'Età

L'età dei pazienti è una variabile cruciale nell'analisi delle comorbilità, poiché molte patologie si manifestano o si aggravano in determinate fasce d'età. Per garantire una rappresentazione equa delle diverse età, il dataset è stato suddiviso in **decadi di nascita**. Le fasce di età considerate sono:

- 1960-1969
- 1970-1979
- 1980-1989
- 1990-1999
- 2000-2009

L'età media dei pazienti è stata mantenuta intorno ai **45 anni**, un'età critica in cui molte comorbilità iniziano a manifestarsi e dove il rischio di malattie croniche è più elevato. Questo bilanciamento ha garantito un equilibrio tra pazienti più giovani e più anziani, evitando che una fascia d'età fosse sovra-rappresentata rispetto alle altre.

Per raggiungere questo obiettivo, sono state utilizzate tecniche di aggregazione e analisi statistica, con l'ausilio della libreria **Pandas**. La distribuzione delle età è stata verificata e, dove necessario, corretta tramite le tecniche di **oversampling** per le classi minoritarie o **undersampling** per le classi sovra-rappresentate, mantenendo così una rappresentanza equa di pazienti in ogni fascia.

Bilanciamento per Sesso

Un altro aspetto cruciale per evitare distorsioni nelle analisi di comorbilità è stato il **bilanciamento per sesso**. Molte malattie hanno impatti differenti su maschi e femmine, ed è importante che il dataset rifletta in modo adeguato entrambe le popolazioni.

Per garantire una distribuzione bilanciata, è stato applicato il **resampling** anche sulla variabile sesso. Quando una classe (maschi o femmine) risultava sottorappresentata, è stato applicato l'**oversampling** per aumentarne la rappresentanza. Allo

stesso modo, per ridurre l'impatto di eventuali squilibri, l'**undersampling** è stato utilizzato sulla classe sovra-rappresentata.

Questo approccio ha garantito che il numero di pazienti maschi e femmine fosse distribuito meglio ma pur mantenendo un grado di rappresentività del dataset iniziale con una leggera predominanza del genere femminile.

4.8 Fase 7: Final Optimization e Analisi Risultati

L'**ottimizzazione finale** rappresenta l'ultimo passaggio nel processo di preparazione del dataset. In questa fase, ci si assicura che il dataset sia completamente bilanciato, demograficamente rappresentativo e privo di errori o incoerenze che potrebbero influenzare negativamente le analisi successive.

Bilanciamento finale

Dopo aver completato le fasi di **data cleaning**, **feature scaling** e **bilanciamento delle variabili chiave**, è stato eseguito un controllo conclusivo per garantire che il dataset rimanesse bilanciato anche nelle variabili, come **età**, **sesso** e **prescrizioni per paziente**. In particolare, si è voluto assicurare che nessuna categoria di pazienti fosse sovra o sotto-rappresentata rispetto alle altre.

- **Età**: Le fasce d'età sono state ulteriormente bilanciate per mantenere una rappresentanza uniforme tra le decadi di nascita già definite nelle fasi precedenti. È stata verificata la distribuzione statistica delle età per evitare che un gruppo specifico dominasse l'analisi.

- **Sesso**: Il bilanciamento tra pazienti maschi e femmine è stato nuovamente controllato, utilizzando le tecniche di **resampling** già menzionate, per garantire che i due sessi fossero rappresentativi del dataset iniziale ma con un bilanciamento migliore.

- **Numero di prescrizioni per paziente**: Anche il numero di prescrizioni è stato oggetto di un'ulteriore analisi, con l'obiettivo di assicurare che i pazienti con poche o troppe prescrizioni non fossero sovra-rappresentati.

Verifica Finale e Coerenza del Dataset

La verifica finale ha previsto un controllo approfondito della coerenza interna del dataset. Questa fase è cruciale per garantire che tutti i dati siano correttamente allineati e non vi siano incongruenze che potrebbero inficiare l'accuratezza dei modelli predittivi o delle analisi di comorbilità.

Le operazioni di verifica includono:

- **Controllo di validità delle relazioni tra le variabili**: Verifica della corretta corrispondenza tra i campi demografici (ad esempio, età e sesso) e le informazioni cliniche (come diagnosi e prescrizioni), per assicurarsi che non vi siano anomalie o errori di associazione.
- **Validazione della completezza del dataset**: Conferma che tutte le colonne chiave siano complete e che le operazioni di eliminazione dei campi nulli abbiano correttamente rimosso le informazioni incomplete o incoerenti.
- **Controllo della qualità del dato ICD9-CM**: Dopo la correzione dei codici anomali, è stata effettuata un'ulteriore revisione dei dati clinici, assicurando che non vi fossero rimasti codici errati o mancanti.

4.8.1 Analisi dei Risultati Pre e Post Bilanciamento

L'analisi dei risultati è un passaggio cruciale per valutare l'efficacia delle operazioni di pulizia, bilanciamento e ottimizzazione del dataset. In questa sezione, andremo ad analizzare i risultati ottenuti dopo l'applicazione delle tecniche di pulizia e bilanciamento dei dati, confrontando il dataset originario da 17 milioni di record con il dataset bilanciato da 40.000 record. Per fare ciò, ci concentreremo su quattro aspetti fondamentali che riflettono le principali caratteristiche demografiche e cliniche: **età**, **codici ICD9-CM**, **numero di prescrizioni per paziente** e **distribuzione per sesso**. Questi aspetti sono cruciali per garantire una rappresentazione equa e bilanciata della popolazione e per evitare distorsioni nei risultati predittivi e nelle analisi di comorbilità.

Di seguito, analizziamo nel dettaglio ciascun aspetto, illustrando l'effetto delle operazioni di bilanciamento e pulizia applicate.

Distribuzione dell'Età

Uno dei fattori fondamentali nell'analisi delle comorbilità è l'età dei pazienti. La maggior parte delle patologie croniche, infatti, tende a manifestarsi con maggiore frequenza in pazienti anziani. Tuttavia, una distribuzione non bilanciata dell'età può distorcere le analisi cliniche, concentrando l'attenzione su fasce d'età particolari e trascurando le altre.

- **Dataset Pre-Bilanciamento**: Nel dataset originario, osserviamo una forte concentrazione di pazienti anziani, soprattutto nella fascia sopra i 60 anni. Questo sbilanciamento riflette la prevalenza di malattie croniche in età avanzata, ma rende il dataset meno rappresentativo dell'intera popolazione clinica.

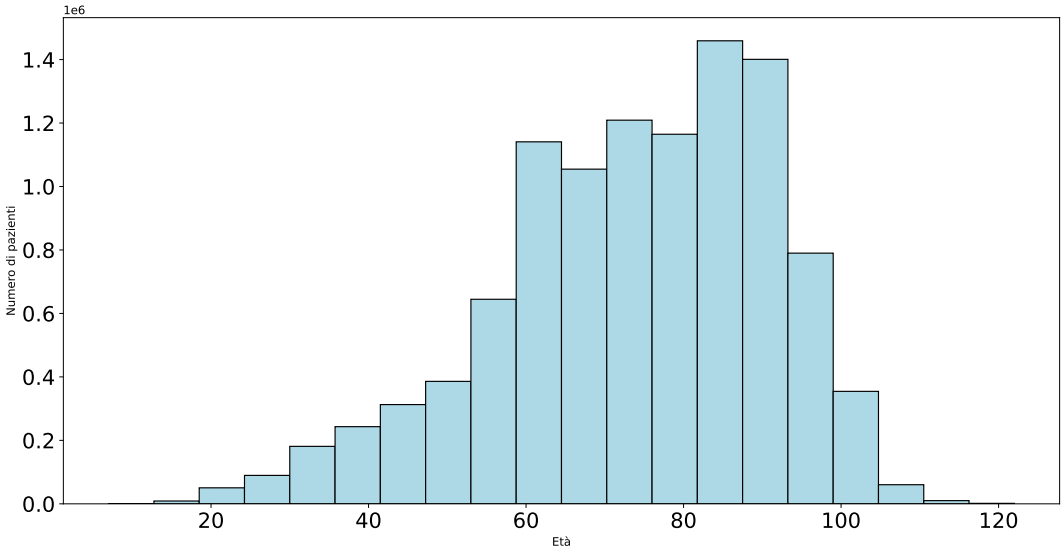

Figura 4.1: Distribuzione età pre-bilanciamento

- **Dataset Post-Bilanciamento**: Dopo l'applicazione del bilanciamento, la distribuzione dell'età è stata resa più uniforme, con una rappresentazione equa delle fasce d'età comprese tra i 30 e i 60 anni. Questo bilanciamento ha migliorato l'accuratezza delle analisi predittive, garantendo che l'algoritmo di machine learning non fosse influenzato dalla predominanza di una singola fascia di età. La distribuzione risulta più bilanciata e rappresentativa.

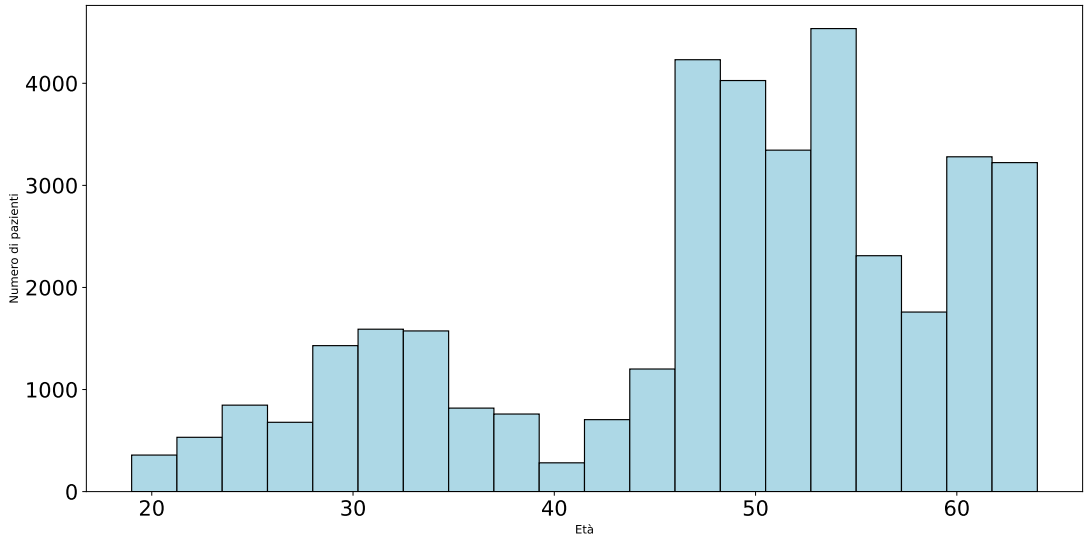

Figura 4.2: Distribuzione età post-bilanciamento

Distribuzione dei Codici ICD9-CM

La classificazione ICD9-CM è essenziale per identificare le malattie e analizzare la loro comorbilità. Un dataset distorto nella distribuzione di questi codici potrebbe favorire alcune patologie a scapito di altre, alterando i risultati dell'analisi delle comorbilità.

- **Dataset Pre-Bilanciamento**: Nel dataset originario, i codici ICD9-CM sono distribuiti in maniera disomogenea, con alcune malattie particolarmente comuni (ad esempio, diabete e ipertensione) che dominano il dataset. Questo squilibrio rischia di rendere meno visibili le relazioni tra malattie meno frequenti.

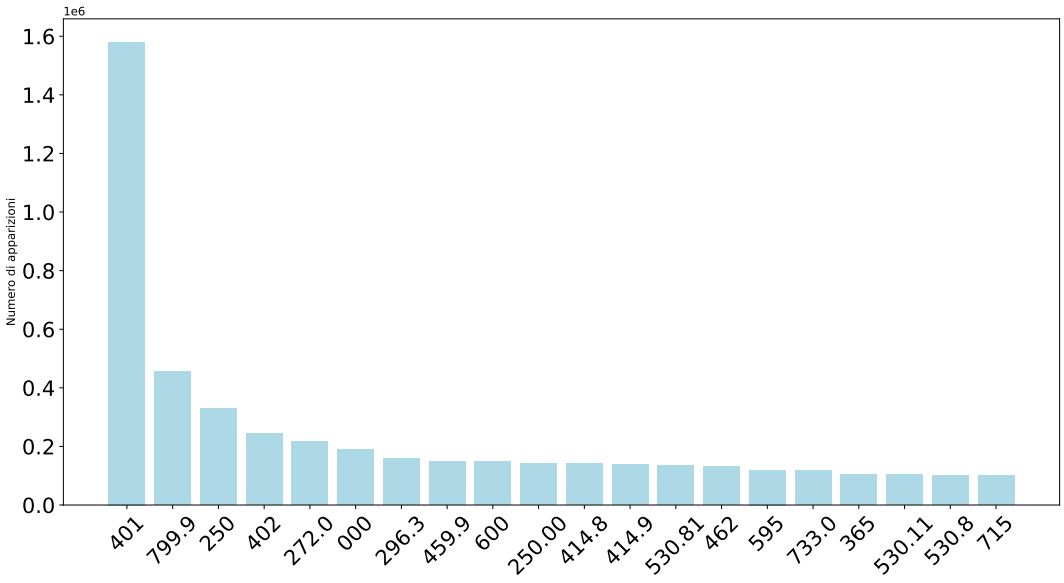

Figura 4.3: Distribuzione codici ICD9-CM pre-bilanciamento

- **Dataset Post-Bilanciamento**: Grazie alle tecniche di bilanciamento, la distribuzione dei codici ICD9-CM risulta più uniforme. Malattie che in precedenza erano sottorappresentate sono state bilanciate, consentendo un'analisi più equa e omogenea delle comorbilità tra le diverse patologie. Questo bilanciamento è cruciale per evitare bias nei modelli predittivi e garantire che tutte le malattie abbiano un peso adeguato nell'analisi.

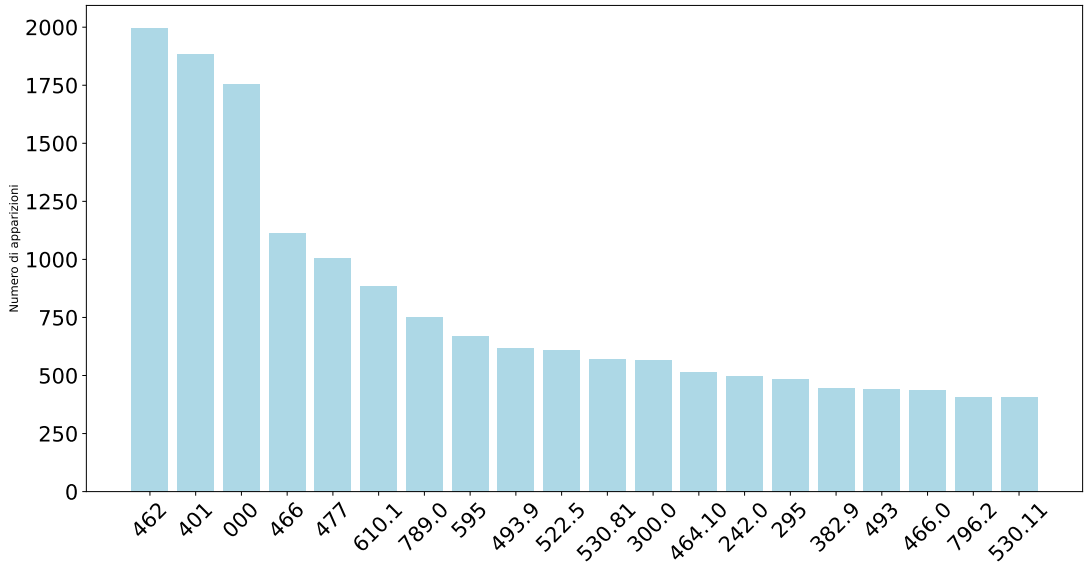

Figura 4.4: Distribuzione codici ICD9-CM post-bilanciamento

Numero di Prescrizioni per Paziente

Il numero di prescrizioni per paziente è un indicatore chiave dell'interazione clinica tra il paziente e il sistema sanitario. Tuttavia, una distribuzione fortemente disomogenea, con alcuni pazienti che ricevono migliaia di prescrizioni, può distorcere l'analisi complessiva.

- **Dataset Pre-Bilanciamento**: Nel dataset originario, si osserva una distribuzione estremamente sbilanciata, con alcuni pazienti che hanno un numero di prescrizioni eccezionalmente alto (fino a 10.000). Questi outlier possono influenzare negativamente le analisi e ridurre la generalizzabilità dei risultati.

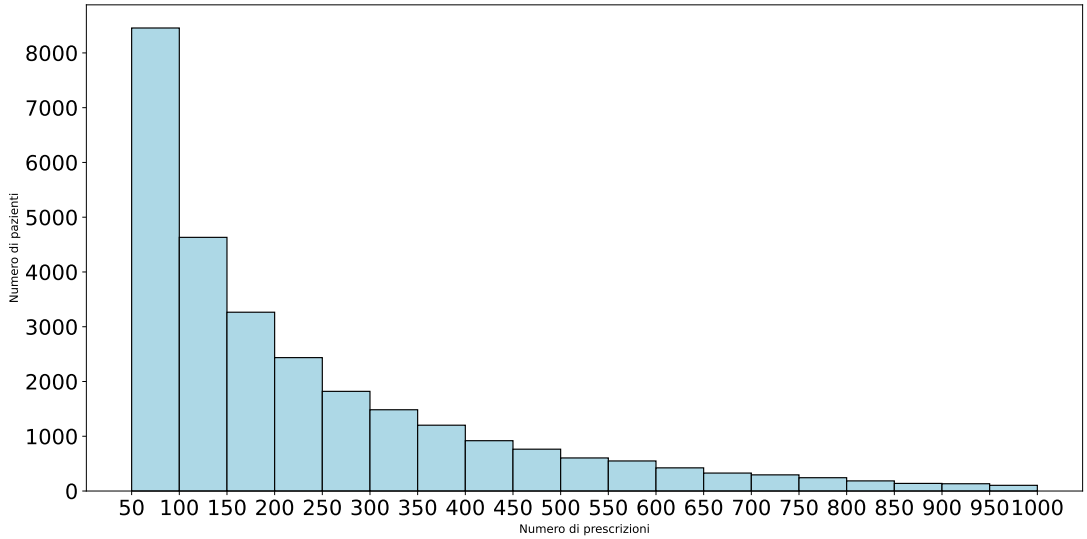

Figura 4.5: Distribuzione Prescrizioni per Paziente pre-bilanciamento

- **Dataset Post-Bilanciamento**: Dopo l'applicazione del bilanciamento, il numero di prescrizioni per paziente è stato limitato tra un minimo di 10 e un massimo di 250. Questo intervallo riflette più accuratamente l'interazione clinica media, eliminando gli outlier e garantendo una maggiore coerenza nelle analisi. Il bilanciamento consente di evitare distorsioni nei modelli predittivi e di migliorare la qualità delle analisi.

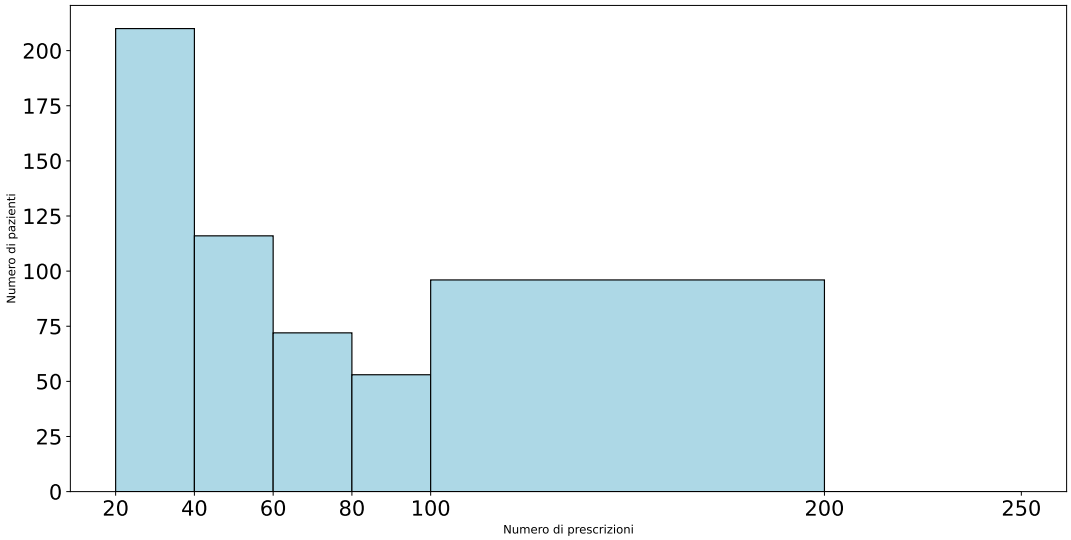

Figura 4.6: Distribuzione Prescrizioni per Paziente post-bilanciamento

Distribuzione del Genere

Infine, la distribuzione del genere è un fattore determinante per garantire che l'analisi delle comorbilità sia rappresentativa di entrambi i generi. Malattie diverse possono avere impatti differenti su uomini e donne, e un dataset sbilanciato potrebbe introdurre bias nei modelli predittivi.

- **Dataset Pre-Bilanciamento**: Nel dataset originale, la distribuzione del genere presenta un leggero sbilanciamento a favore dei pazienti maschi. Questo squilibrio potrebbe influenzare l'accuratezza delle analisi, soprattutto nelle malattie con una prevalenza diversa tra i sessi.

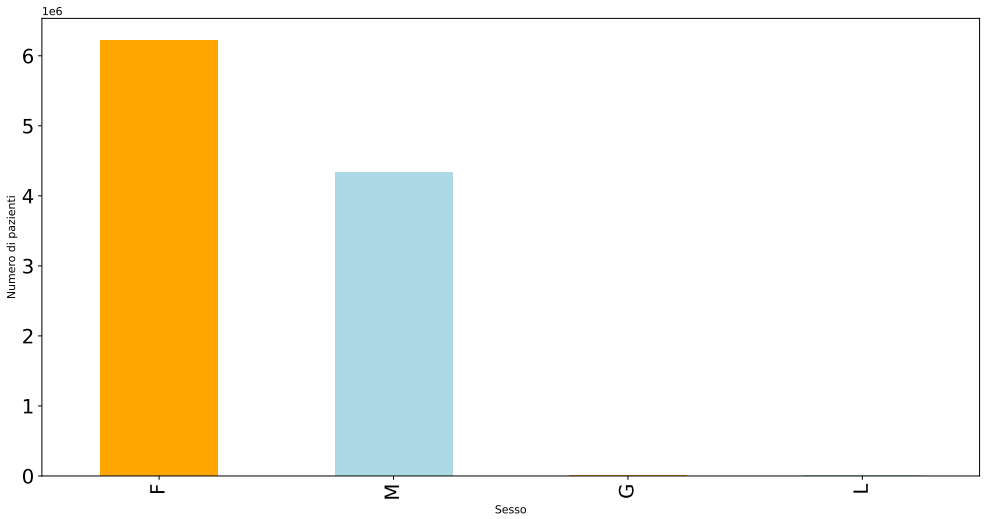

Figura 4.7: Distribuzione genere pre-bilanciamento

- **Dataset Post-Bilanciamento**: Il bilanciamento del dataset ha corretto questa disuguaglianza, garantendo una distribuzione equa tra pazienti maschi e femmine. Ciò assicura che le analisi delle comorbilità siano applicabili a entrambi i sessi e che i modelli di machine learning possano fare previsioni più accurate.

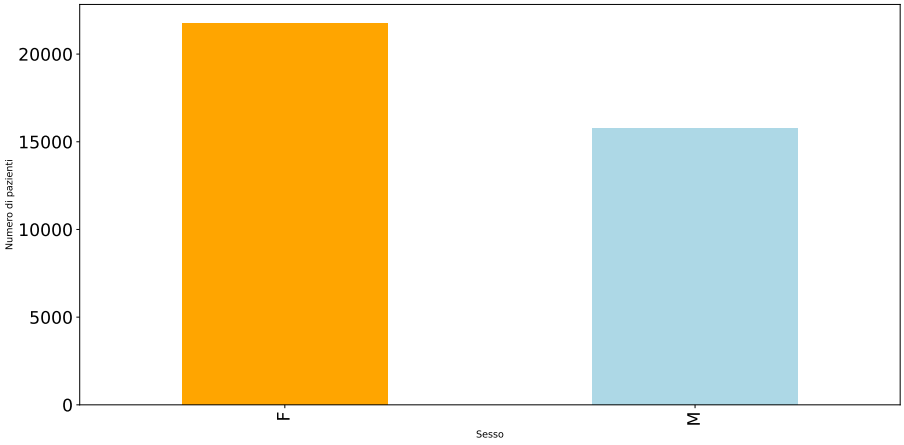

Figura 4.8: Distribuzione genere post-bilanciamento

4.9 Conclusioni

L'analisi dei risultati pre e post bilanciamento evidenzia chiaramente il valore delle operazioni di pulizia e bilanciamento applicate al dataset. Il dataset bilanciato da 40.000 record non solo riflette in modo più accurato la popolazione clinica, ma garantisce anche che le analisi delle comorbilità e i modelli di machine learning possano produrre risultati più robusti, accurati e generalizzabili. Le operazioni di bilanciamento hanno migliorato notevolmente la qualità e l'affidabilità del dataset.

CAPITOLO 5

Social Network Analysis (SNA) e Visual Analytics (VA) per l'Analisi della Comorbidità

5.1 Introduzione

La **Social Network Analysis (SNA)** è una tecnica ampiamente utilizzata per studiare relazioni complesse all'interno di sistemi rappresentati come grafi. In un grafo, le entità sono rappresentate come nodi, mentre le relazioni tra queste entità sono gli archi che li connettono. Nel contesto clinico, e in particolare nello studio della comorbidità, la SNA si rivela una metodologia particolarmente potente per analizzare le interconnessioni tra patologie, pazienti e trattamenti medici.

In ComorGraph, la SNA è implementata per visualizzare e analizzare la rete delle malattie comorbide. Questa rete è composta da nodi che rappresentano le malattie, i pazienti e le prescrizioni, mentre gli archi tra i nodi indicano le relazioni di comorbidità tra malattie o trattamenti condivisi. Questo tipo di rappresentazione consente di individuare pattern che altrimenti rimarrebbero nascosti con una struttura di dati tradizionale. Malattie che si manifestano spesso insieme possono essere identificate come nodi centrali nella rete, mentre altre patologie possono fungere da "ponte" tra gruppi distinti di malattie.

L'uso della SNA nella piattaforma ComorGraph è fondamentale per raggiungere tre obiettivi principali:

1. **Identificazione di malattie centrali o "hub"**: Le malattie che svolgono un ruolo cruciale nella rete di comorbidità, cioè quelle che sono frequentemente correlate ad altre patologie, vengono evidenziate. Questo è particolarmente utile per le malattie croniche, come il diabete o l'ipertensione, che spesso fungono da "snodo" per molte altre condizioni.

2. **Scoperta di malattie "ponte"**: Alcune malattie possono connettere diverse aree della rete, fungendo da legame tra patologie che sembrano altrimenti non correlate. Identificare questi nodi "ponte" permette di comprendere meglio le dinamiche di propagazione delle condizioni cliniche tra diversi gruppi di patologie.

3. **Previsione di pattern di malattie future**: Analizzando i collegamenti tra malattie note, la SNA consente di fare predizioni su potenziali sviluppi futuri della salute di un paziente, suggerendo possibili complicazioni o nuove diagnosi sulla base delle comorbidità osservate.

Grazie a questi tre obiettivi, ComorGraph permette una visualizzazione chiara e interattiva delle reti di comorbidità, supportando i medici nella gestione delle malattie dei pazienti in modo più consapevole e basato sui dati.

5.2 Degree Centrality

La **Degree Centrality** è una metrica chiave nella Social Network Analysis (SNA) per valutare l'importanza di un nodo all'interno di una rete in base al numero di connessioni dirette che possiede. Nella piattaforma ComorGraph, questa metrica viene utilizzata per identificare le malattie che hanno più associazioni con altre patologie all'interno della rete di comorbidità. Tuttavia, una particolarità dell'approccio di ComorGraph è che non solo vengono contate le connessioni tra malattie, ma viene anche preso in considerazione un **counter** sulla relazione tra paziente e malattia, che influisce sui calcoli della centralità.

Contesto della Relazione `DIAGNOSTICATO_CON`

Per ridurre la complessità delle relazioni nel grafo e minimizzare il numero di connessioni dirette tra i pazienti e le malattie, la relazione `DIAGNOSTICATO_CON` tra un paziente e una malattia include un contatore. Questo contatore tiene traccia del numero di volte in cui un paziente ha ricevuto una diagnosi specifica. Pertanto, quando si calcola la degree centrality, il sistema non si limita a contare le connessioni binarie tra pazienti e malattie, ma **somma i valori del contatore**, pesando maggiormente le malattie con una frequenza maggiore di diagnosi ripetute.

Applicazione sulla piattaforma

In ComorGraph, la **Degree Centrality** viene calcolata per ciascuna malattia sommando il numero totale di connessioni ponderate (in base al contatore `DIAGNOSTICATO_CON`) tra una specifica malattia e i pazienti. Questo metodo garantisce una misurazione più accurata dell'importanza di una malattia nella rete di comorbidità, poiché malattie con diagnosi frequenti avranno un peso maggiore. Questo approccio è particolarmente utile per mettere in luce malattie croniche o ricorrenti, che giocano un ruolo cruciale nella salute complessiva del paziente.

Un'altra caratteristica chiave di ComorGraph è che i risultati della degree centrality vengono **raggruppati per "gruppi di malattie"**. Questi gruppi sono basati sui 20 gruppi identificati nello studio del dottor Cavallo, che categorizzano le malattie secondo criteri clinici condivisi. La piattaforma visualizza questi risultati nella sezione dashboard, dove gli utenti possono osservare il **grado di centralità aggregato** per ciascun gruppo di malattie. La visualizzazione avviene attraverso un grafico interattivo, in cui i gruppi con la degree centrality più alta emergono chiaramente, facilitando così l'interpretazione dei pattern di comorbidità.

Importanza clinica della Degree Centrality:

Questa metrica è particolarmente importante per individuare le malattie centrali che, se trattate in modo efficace, potrebbero avere un impatto positivo su altre malattie collegate. Malattie con un **alto degree centrality**, come il diabete, tendono ad essere collegate a numerose altre patologie croniche. Il fatto che queste malattie siano identificate come "hub" nella rete suggerisce che la gestione di queste condizioni chiave potrebbe ridurre l'aggravarsi di altre patologie.

Inoltre, visualizzare i risultati raggruppati per categorie cliniche migliora ulteriormente l'analisi e la comprensione delle reti di comorbidità. Ad esempio, un gruppo di malattie con alta degree centrality potrebbe indicare condizioni fortemente interconnesse che richiedono un approccio terapeutico combinato.

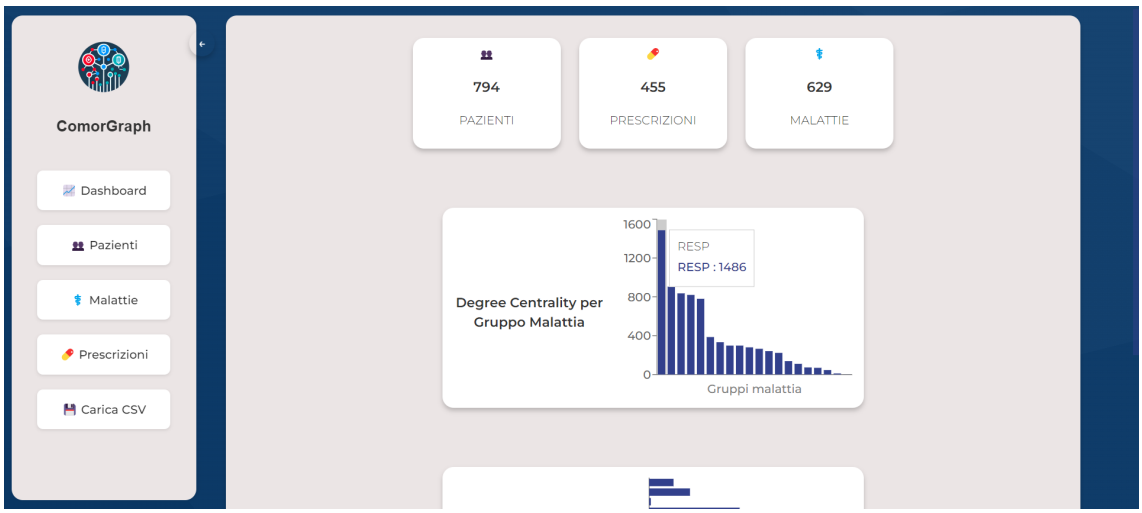

Figura 5.1: Degree Centrality su ComorGraph

5.3 Betweenness Centrality

La **Betweenness Centrality** è una delle principali metriche utilizzate nella **Social Network Analysis (SNA)** per identificare i nodi che fungono da intermediari nel grafo. Essa misura quanto un nodo sia cruciale per la connessione di altri nodi, agendo come "ponte" nei percorsi più brevi tra di essi. In sostanza, individua quelle entità che facilitano il flusso di informazioni o influenze tra diverse aree di una rete.

Interpretazione in un contesto di comorbilità Nel campo delle **reti di comorbilità**, questa metrica è fondamentale per comprendere quali malattie collegano cluster distinti di patologie. Le malattie con alta betweenness centrality sono quelle che fungono da "ponte" tra gruppi di malattie che altrimenti sarebbero separate. Questo implica che tali malattie potrebbero avere un ruolo centrale nella progressione clinica di condizioni multiple e che, intervenendo su di esse, si potrebbe influenzare positivamente l'evoluzione di più patologie associate.

In termini clinici, una malattia con alta betweenness potrebbe essere cruciale per il collegamento tra malattie croniche complesse e altre condizioni. Ad esempio, una malattia come l'obesità potrebbe fungere da connessione tra il diabete e le malattie cardiovascolari, suggerendo che il trattamento di questa condizione centrale potrebbe avere un impatto su entrambi i gruppi.

Calcolo nel contesto della piattaforma ComorGraph Nella piattaforma **ComorGraph**, il calcolo della **Betweenness Centrality** viene effettuato proiettando il grafo delle sole malattie, la metrica si basa sulle relazioni **'ASSOCIATA_A'**. Queste relazioni rappresentano i casi di comorbidità, ossia la co-presenza di due malattie in più pazienti.

In questo contesto, viene eseguita una **proiezione del grafo** in cui le malattie sono i nodi e le relazioni **ASSOCIATA_A** sono archi non direzionati. Si considera inoltre il valore **'count'** come peso degli archi, per riflettere la frequenza con cui queste due malattie si presentano insieme nei pazienti.

Il risultato è una visualizzazione chiara di quelle malattie che fungono da "ponti" cruciali nel grafo delle comorbidità. Queste malattie non sono semplicemente colle-

gate a molte altre patologie, ma collegano diversi cluster di malattie, facilitando la propagazione delle condizioni cliniche all'interno della rete.

Visualizzazione e Risultati sulla Piattaforma Nella piattaforma **ComorGraph**, i nodi che mostrano una **Betweenness Centrality** elevata vengono immediatamente riconosciuti grazie alla loro dimensione. I nodi con un valore maggiore vengono rappresentati con dimensioni più grandi rispetto agli altri, fornendo così una visione intuitiva delle malattie più strategicamente posizionate nella rete. Se si vuole approfondire sul valore di ciasun nodo è possibile visualizzarlo nella sezione dettagli.

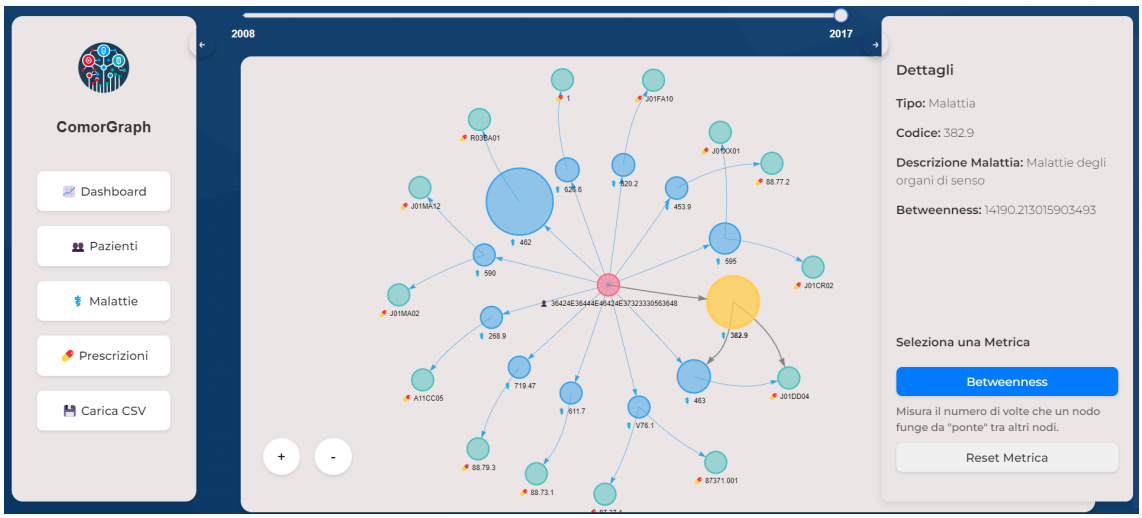

Figura 5.2: Betweenness Centrality su ComorGraph

Questo tipo di rappresentazione aiuta i medici e i ricercatori a identificare visivamente le malattie che fungono da ponti tra diversi cluster di patologie. Tali malattie, se trattate in modo efficace, potrebbero influenzare l'intero network di comorbidità e migliorare la gestione complessiva delle condizioni cliniche dei pazienti.

5.4 Closeness Centrality

La **Closeness Centrality** è una metrica fondamentale nella **Social Network Analysis (SNA)**, che misura la vicinanza di un nodo rispetto a tutti gli altri nodi all'interno del grafo. In altre parole, questa metrica valuta quanto velocemente un nodo può raggiungere tutti gli altri nodi della rete, e quindi quanto è centrale nel network.

Interpretazione in un contesto di comorbilità In una rete di comorbidità, la **Closeness Centrality** è utile per identificare quelle malattie che sono "vicine" ad altre patologie, ovvero quelle condizioni che possono rapidamente influenzare o essere influenzate da altre malattie. Le malattie con un'alta closeness centrality tendono a trovarsi in posizioni strategiche della rete, con percorsi brevi verso molte altre patologie.

Clinicamente, una malattia con alta **Closeness Centrality** può rappresentare una condizione che, se trattata, potrebbe avere effetti diretti e rapidi su numerose altre patologie. Questo la rende cruciale per la gestione della salute del paziente, poiché intervenendo tempestivamente su di essa si può influenzare il decorso di molte altre malattie associate.

Calcolo nel contesto della piattaforma ComorGraph Nella piattaforma **ComorGraph**, la **Closeness Centrality** viene calcolata proiettando il grafo delle malattie, utilizzando le sole relazioni **'ASSOCIATA_A'**, che rappresentano la comorbidità tra le patologie. Analogamente alla **Betweenness Centrality**, si utilizza una proiezione del grafo che considera solo le malattie come nodi e le loro connessioni dirette come archi. Il valore **'count'**, che rappresenta la frequenza delle comorbidità tra due malattie, non viene considerato come peso in questo calcolo, poiché l'obiettivo è valutare la vicinanza puramente strutturale delle malattie all'interno della rete.

Il risultato di questo calcolo evidenzia quelle malattie che possono essere raggiunte più velocemente da tutte le altre, suggerendo che queste patologie giocano un ruolo chiave nella diffusione o nella propagazione delle condizioni all'interno del network.

Visualizzazione e Risultati sulla Piattaforma Nella piattaforma **ComorGraph**, le malattie con elevata **Closeness Centrality** vengono visualizzate con nodi di dimensioni maggiori, simili alle altre metriche, per facilitare l'identificazione di quelle malattie che sono "centrali" rispetto a molte altre. Il criterio visivo è immediato: le malattie più vicine ad altre patologie verranno visualizzate come nodi centrali e prominenti all'interno del grafo.

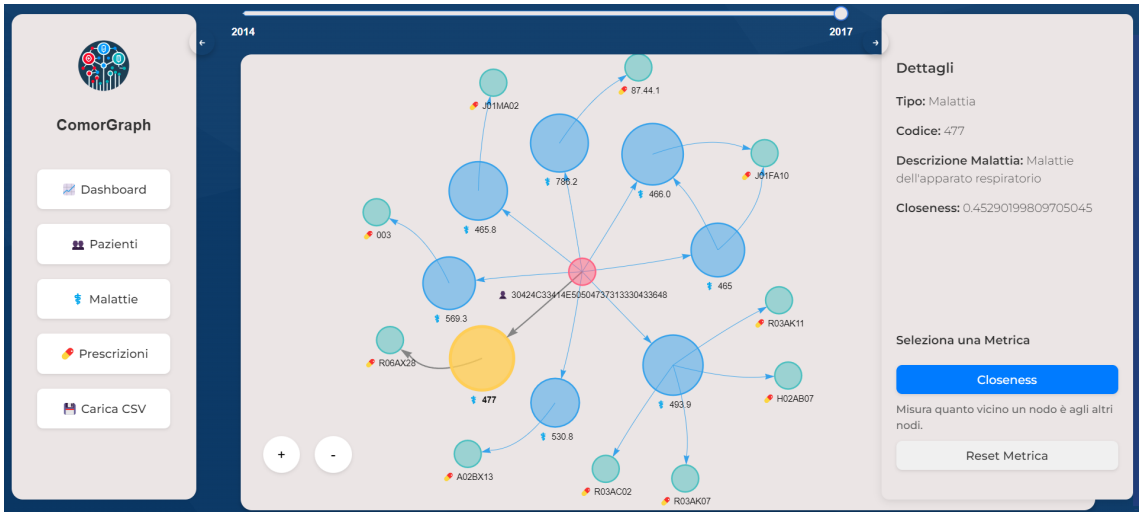

Figura 5.3: Closeness Centrality su ComorGraph

Questa rappresentazione permette ai clinici di comprendere rapidamente quali malattie devono essere monitorate da vicino per la loro capacità di influenzare rapidamente altre condizioni, offrendo così preziosi spunti per la pianificazione terapeutica.

5.5 PageRank

Il **PageRank** è una metrica ampiamente conosciuta per il suo utilizzo originario nell'algoritmo di ranking di Google. La sua funzione è assegnare un valore di "importanza" a ciascun nodo del grafo in base non solo al numero di connessioni che il nodo ha, ma anche all'importanza dei nodi a cui è collegato. Questo consente di stabilire una gerarchia tra i nodi, dove i nodi con le connessioni più rilevanti (cioè, connessi a nodi di grande influenza) ricevono un punteggio più elevato.

Interpretazione in un contesto di comorbilità Nel contesto della comorbilità, il **PageRank** permette di identificare le malattie non solo per il numero di connessioni che hanno con altre patologie, ma anche per il loro legame con malattie influenti. Questo significa che una malattia con un PageRank elevato potrebbe essere una condizione meno comune, ma fortemente associata a patologie centrali e di grande rilevanza clinica.

In un'analisi di comorbilità, il **PageRank** aiuta a individuare quelle malattie che, sebbene non abbiano molte connessioni dirette, sono strettamente legate a patologie chiave, e perciò assumono un'importanza notevole nella rete. Malattie con alto PageRank potrebbero rivelarsi cruciali per comprendere come certe condizioni si diffondono o influenzano altre patologie, e trattarle potrebbe avere un impatto significativo sulla salute generale del paziente.

Calcolo nel contesto della piattaforma ComorGraph Sulla piattaforma **ComorGraph**, il **PageRank** viene calcolato considerando solo i nodi di tipo **Malattia** e le loro relazioni **'ASSOCIATA_A'**, senza utilizzare il peso **'count'** delle relazioni. Ciò significa che il **PageRank** valuta l'importanza di una malattia non solo in base al numero di connessioni con altre malattie, ma anche in base all'importanza delle malattie con cui è collegata.

In questo contesto, malattie che possono non avere molte connessioni dirette, ma che sono collegate a malattie altamente centrali o influenti, otterranno comunque un valore di **PageRank** elevato. Questo permette di identificare patologie che, pur

non essendo immediatamente evidenti come centrali, giocano un ruolo cruciale nel mantenere connessa la rete delle comorbidità.

Visualizzazione e Risultati sulla Piattaforma Nella piattaforma **ComorGraph**, le malattie con un elevato PageRank vengono evidenziate in modo prominente all'interno del grafo. I nodi più influenti, quelli con il più alto punteggio di PageRank, saranno visualizzati con dimensioni maggiori, rendendo evidente all'utente quali sono le malattie di maggiore impatto all'interno della rete di comorbidità.

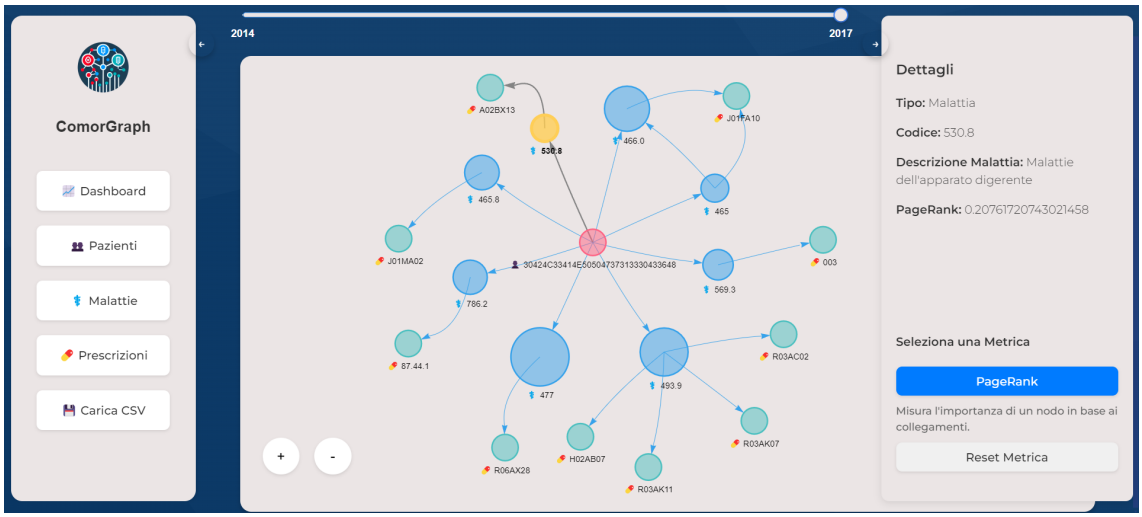

Figura 5.4: PageRank su ComorGraph

Questa rappresentazione permette ai medici e ai ricercatori di individuare malattie "nascoste" ma influenti, che potrebbero non essere immediatamente evidenti attraverso altre metriche. Tale visualizzazione facilita la comprensione dell'importanza relativa delle patologie nel contesto clinico e fornisce spunti preziosi per la gestione delle comorbidità.

5.6 K-Core

Il **K-Core** è una metrica che identifica gruppi di nodi fortemente connessi tra loro. Un **K-Core** è un sottografo in cui ogni nodo ha almeno **k** connessioni con altri nodi all'interno dello stesso sottografo. Questo concetto è particolarmente utile quando si vogliono individuare "cluster" o comunità ben definite all'interno di una rete. Nel contesto della comorbilità, il **K-Core** permette di identificare gruppi di malattie che tendono a coesistere frequentemente in gruppi specifici di pazienti, formando un "nucleo" di comorbidità.

Interpretazione in un contesto di comorbilità

L'applicazione del **K-Core** alla rete di comorbilità permette di individuare gruppi di malattie che presentano una forte coesione interna, vale a dire malattie che tendono a manifestarsi frequentemente insieme nei pazienti. Questo tipo di cluster può essere particolarmente utile per identificare sindromi complesse o condizioni cliniche multi-sistemiche, in cui un gruppo di malattie coesiste con alta frequenza. Questi cluster possono fornire informazioni preziose ai medici per identificare pattern di malattie che potrebbero suggerire interventi terapeutici congiunti o approcci di prevenzione mirata.

Ad esempio, un gruppo di malattie che forma un **K-Core** potrebbe essere composto da patologie come il diabete, l'ipertensione e le malattie cardiovascolari, che spesso compaiono insieme nei pazienti con comorbidità. L'identificazione di questi gruppi facilita l'analisi delle interazioni tra malattie e la pianificazione clinica.

Calcolo nel contesto della piattaforma ComorGraph

Nella piattaforma **ComorGraph**, il **K-Core** viene calcolato proiettando un grafo che considera solo i nodi **Malattia** e le loro relazioni **ASSOCIATA_A**. Ogni malattia è collegata alle altre in base alla presenza di comorbidità tra i pazienti, e l'algoritmo del **K-Core** identifica i gruppi di malattie che formano nuclei fortemente connessi.

In questo modo, il **K-Core** evidenzia cluster di malattie che possono fornire un'indicazione chiara su quali patologie tendono a manifestarsi insieme in modo coeso. Una volta calcolato il **K-Core**, i risultati sono rappresentati visivamente sulla

piattaforma, con i nodi ridimensionati in base al loro grado di appartenenza ai core più elevati. Sebbene questa rappresentazione possa sembrare semplificata rispetto a un'analisi più dettagliata delle connessioni interne ai cluster, ha il vantaggio di essere immediata e intuitiva.

L'uso di dimensioni diverse per i nodi facilita una lettura visiva rapida, consentendo agli utenti della piattaforma di identificare immediatamente le malattie più rilevanti all'interno di ciascun cluster di comorbidità. Questo approccio semplifica la comprensione anche per utenti non tecnici, come medici e professionisti sanitari, che possono così concentrarsi sugli aspetti clinici piuttosto che su una rappresentazione troppo complessa. I nodi più grandi indicano malattie che giocano un ruolo centrale nei core di comorbidità più elevati, mentre i nodi più piccoli indicano malattie meno connesse.

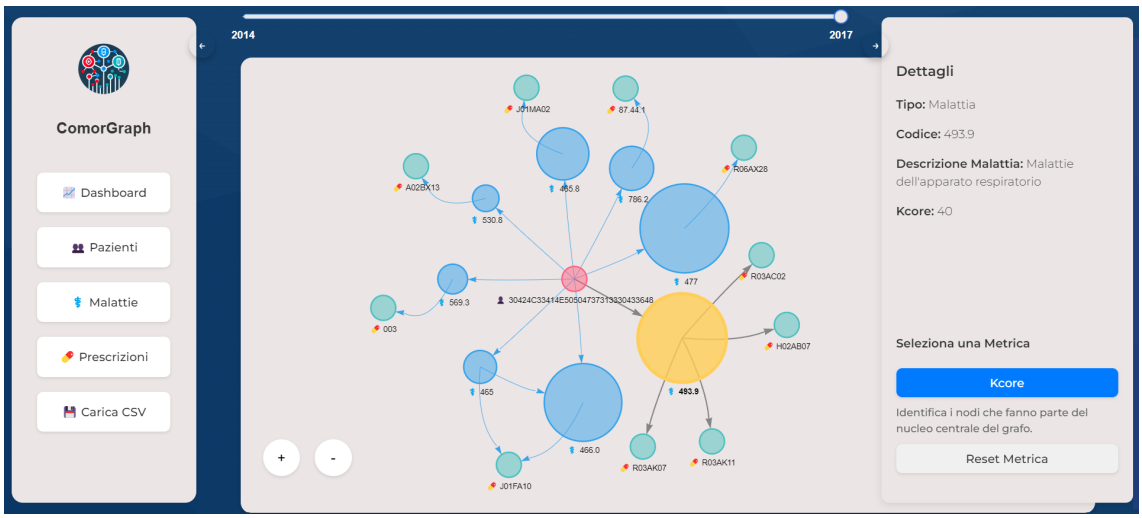

Figura 5.5: K-Core su ComorGraph

Questa rappresentazione visuale immediata e intuitiva, con le dimensioni dei nodi che riflettono l'importanza delle malattie nei vari **K-Core**, permette una rapida interpretazione dei risultati senza rinunciare alla precisione. Le foto che illustrano la rappresentazione del **K-Core** sulla piattaforma possono essere allegate per mostrare come questo approccio renda accessibili i cluster di comorbidità, fornendo così un supporto visivo essenziale per i medici.

CAPITOLO 6

Conclusioni e Sviluppi Futuri

6.1 Conclusioni

Le conclusioni di questo lavoro di tesi evidenziano il valore della piattaforma **ComorGraph** come strumento innovativo per l'analisi delle comorbidità nel contesto sanitario. L'integrazione di tecnologie avanzate come **Neo4j**, **Python**, e **React** ha permesso di costruire una piattaforma capace di visualizzare e analizzare relazioni complesse tra malattie, offrendo supporto ai professionisti nella gestione delle condizioni multiple dei pazienti.

La rappresentazione dei dati tramite grafi e l'adozione di tecniche di **Social Network Analysis (SNA)** hanno giocato un ruolo chiave, consentendo di individuare malattie "hub" e di identificare nodi che fungono da "ponte" tra diversi gruppi di patologie. Queste tecniche hanno migliorato la comprensione delle reti di comorbidità, con un impatto diretto sulla capacità di gestione clinica e diagnosi predittiva. La combinazione di queste metriche con strumenti di **Visual Analytics** ha reso il processo di interpretazione più immediato e intuitivo, fornendo un quadro chiaro delle interconnessioni tra le malattie.

Un punto di forza ulteriore è stato il riconoscimento da parte del dottor Pierpaolo Cavallo, che ha espresso soddisfazione per la piattaforma e per la sua efficacia nel

rispondere agli obiettivi di ricerca posti. ComorGraph non solo ha raggiunto le aspettative iniziali, ma ha suscitato l'interesse per futuri sviluppi e potenziali studi di approfondimento.

Tuttavia, lo sviluppo della piattaforma non è stato privo di sfide. La gestione dei dati clinici, in particolare quelli storici (dal 1900 al 2004), ha richiesto un complesso lavoro di pulizia e normalizzazione. La qualità variabile e la natura disomogenea dei dati hanno richiesto sforzi significativi per garantire la coerenza e l'accuratezza delle analisi. Questo è stato un passaggio cruciale per assicurare la solidità della piattaforma e la validità delle sue previsioni.

Un altro aspetto innovativo è rappresentato dall'integrazione con moduli di intelligenza artificiale basati su **Heterogeneous Graph Neural Networks (HeteroGNN)**, che hanno introdotto capacità predittive per anticipare l'insorgenza di malattie. Sebbene non sia il core della piattaforma, questo modulo ha dimostrato un grande potenziale nell'offrire nuove prospettive per la medicina predittiva.

In conclusione, **ComorGraph** rappresenta un significativo avanzamento nel campo dell'analisi delle comorbidità. Le sue capacità di analisi di rete e di previsione offrono strumenti fondamentali per migliorare l'efficacia dei trattamenti clinici e per supportare decisioni mediche informate. I futuri sviluppi della piattaforma, insieme a ulteriori studi sull'integrazione di modelli predittivi avanzati, potranno portare a nuovi importanti risultati nel campo della medicina personalizzata e predittiva.

6.2 Sviluppi Futuri

Nonostante **ComorGraph** rappresenti già un avanzamento significativo nello studio delle comorbidità grazie alla sua capacità di analizzare reti complesse di malattie, ci sono alcuni sviluppi futuri che potrebbero migliorare ulteriormente la piattaforma, offrendo ai professionisti sanitari strumenti ancora più potenti e specifici. Uno dei principali obiettivi futuri è l'integrazione del modulo di **intelligenza artificiale** direttamente all'interno della piattaforma, permettendo così un'analisi predittiva più precisa ed efficiente. In questo modo, ComorGraph potrà evolversi in uno strumento completo che non solo consente di visualizzare e analizzare le reti di comorbidità, ma anche di anticipare potenziali sviluppi patologici per ogni singolo paziente.

Un secondo aspetto fondamentale che sarà esplorato è l'implementazione di un modello basato su **Temporal Graph Neural Networks (TemporalGNN)**. L'introduzione della dimensione temporale permetterà di studiare i **pattern temporali**, le periodicità e la frequenza con cui determinate malattie si manifestano nei pazienti. Questo potrebbe offrire nuove prospettive nel comprendere la progressione delle patologie e nel predire l'evoluzione della salute di un paziente con un livello di dettaglio molto più alto.

Il terzo potenziale sviluppo è la creazione di un **null model** della rete di comorbidità. Un null model è una rete casuale che conserva alcune proprietà strutturali della rete reale, ma rimuove altre proprietà, consentendo di eseguire **analisi comparative** per identificare quali caratteristiche emergono naturalmente dalla struttura del grafo e quali sono significative rispetto alla casualità. Questo strumento potrebbe fornire ulteriori insight sulle dinamiche delle malattie nelle reti di comorbidità.

Infine, si prevede di includere funzionalità per **evidenziare relazioni ricorrenti** tra malattie. Queste informazioni potranno essere utilizzate per avvisare gli utenti quando malattie che tendono a verificarsi insieme appaiono nel grafico del paziente, migliorando la gestione della diagnosi e prevenendo potenziali complicazioni future.

Questi sviluppi rappresentano opportunità per migliorare la piattaforma, pur mantenendo la sua solidità attuale, e per offrire funzionalità avanzate che potrebbero portare a nuove scoperte e ottimizzazioni nel campo della medicina predittiva e della gestione clinica delle comorbidità.

Bibliografia

[1] J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury, and M. Roland, "Defining comorbidity: Implications for understanding health and health services," *The Annals of Family Medicine*, vol. 7, no. 4, pp. 357–363, 2009. [Online]. Available: https://www.annfammed.org/content/7/4/357 (Citato alle pagine 5 e 6)

[2] R. G. V. M. Prosperina, Ed., *Challenges in Social Network Research: Methods and Applications*. Springer, 2020. [Online]. Available: https://link.springer.com/book/10.1007/978-3-030-31463-7 (Citato a pagina 5)

[3] R. Brown and E. Thorsteinsson, *Comorbidity: What Is It and Why Is It Important?* Cham: Springer International Publishing, 2020, pp. 1–22. [Online]. Available: https://doi.org/10.1007/978-3-030-32545-9_1 (Citato a pagina 6)

[4] D. Chambers, P. Wilson, C. Thompson, and M. Harden, "Social network analysis in healthcare settings: A systematic scoping review," *PLOS ONE*, vol. 7, no. 8, p. e41911, 2012. [Online]. Available: https://doi.org/10.1371/journal.pone.0041911 (Citato a pagina 6)

[5] G. Giordano, M. De Santis, S. Pagano, G. Ragozini, M. P. Vitale, and P. Cavallo, *Association Rules and Network Analysis for Exploring Comorbidity Patterns in Health Systems*. Cham: Springer International Publishing, 2020, pp. 63–78. [Online].

Available: https://doi.org/10.1007/978-3-030-31463-7_5 (Citato alle pagine 6, 22 e 24)

[6] R. M. Payton J. Jones and R. J. McNally, "Bridge centrality: A network approach to understanding comorbidity," *Multivariate Behavioral Research*, vol. 56, no. 2, pp. 353–367, 2021, pMID: 31179765. [Online]. Available: https://doi.org/10.1080/00273171.2019.1614898 (Citato a pagina 7)

[7] C.-W. Huang, R. Lu, U. Iqbal, S.-H. Lin, P. A. A. Nguyen, H.-C. Yang, C.-F. Wang, J. Li, K.-L. Ma, Y.-C. J. Li, and W.-S. Jian, "A richly interactive exploratory data analysis and visualization tool using electronic medical records," *BMC Medical Informatics and Decision Making*, vol. 15, no. 1, p. 92, Nov 2015. [Online]. Available: https://doi.org/10.1186/s12911-015-0218-7 (Citato a pagina 7)

[8] N. Rostamzadeh, S. S. Abdullah, and K. Sedig, "Visual analytics for electronic health records: A review," *Informatics*, vol. 8, no. 1, 2021. [Online]. Available: https://www.mdpi.com/2227-9709/8/1/12 (Citato a pagina 7)

[9] H. Lu and S. Uddin, "Embedding-based link predictions to explore latent comorbidity of chronic diseases," *Health Information Science and Systems*, vol. 11, no. 1, p. 2, 2022. [Online]. Available: https://doi.org/10.1007/s13755-022-00206-7 (Citato a pagina 8)

[10] R. J. Woodman, B. Koczwara, and A. A. Mangoni, "Applying precision medicine principles to the management of multimorbidity: the utility of comorbidity networks, graph machine learning, and knowledge graphs," *Frontiers in Medicine*, vol. 10, p. 1302844, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmed.2023.1302844/full (Citato a pagina 8)

[11] S. Biswas, K. D. Chaudhuri, P. Mitra, and K. S. Rao, "Relation predictions in comorbid disease centric knowledge graph using heterogeneous gnn models," in *Bioinformatics and Biomedical Engineering*, I. Rojas, O. Valenzuela, F. Rojas Ruiz, L. J. Herrera, and F. Ortuño, Eds. Cham: Springer Nature Switzerland, 2023, pp. 343–356. (Citato a pagina 9)

[12] R. Bing, G. Yuan, M. Zhu, F. Meng, H. Ma, and S. Qiao, "Heterogeneous graph neural networks analysis: a survey of techniques, evaluations and applications," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8003–8042, 2023. [Online]. Available: https://doi.org/10.1007/s10462-022-10375-2 (Citato a pagina 10)

[13] N. Shahid, T. Rappon, and W. Berta, "Applications of artificial neural networks in health care organizational decision-making: A scoping review," *PLOS ONE*, vol. 14, no. 2, p. e0212356, 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0212356 (Citato alle pagine 10 e 11)

[14] P. Cavallo, S. Pagano, M. De Santis, and E. Capobianco, "General practitioners records are epidemiological predictors of comorbidities: An analytical cross-sectional 10-year retrospective study," *Journal of Clinical Medicine*, vol. 7, no. 8, p. 184, 2018. [Online]. Available: https://doi.org/10.3390/jcm7080184 (Citato alle pagine 11, 12 e 13)

[15] A. O. Adeniyi, C. A. Okolo, T. Olorunsogo, and O. Babawarun, "Leveraging big data and analytics for enhanced public health decision-making: A global review," *GSC Advanced Research and Reviews*, vol. 18, no. 2, pp. 450–456, 2024. [Online]. Available: https://doi.org/10.30574/gscarr.2024.18.2.0078 (Citato a pagina 13)

[16] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Machine Learning*, vol. 113, no. 7, pp. 4845–4901, 2024. [Online]. Available: https://doi.org/10.1007/s10994-022-06268-8 (Citato a pagina 43)