MentalHealthAnalytics

Analisi dei disturbi mentali attraverso tecniche di data analysis, machine learning e modellazione predittiva.

Autore
Annarita Bruno, 766402, a.bruno113@studenti.uniba.it

< https://github.com/annabr98/consegnaICON2526.git>

AA 2025-2026

INDICE

INTRODUZIONE

Negli ultimi anni la salute mentale ha assunto un ruolo sempre più centrale nel dibattito scientifico e nelle politiche sanitarie, dopo un lungo periodo di relativa sottovalutazione. Il presente lavoro analizza alcune delle principali patologie mentali, tra cui la schizofrenia, i disturbi depressivi, i disturbi d'ansia, i disturbi bipolari e i disturbi del comportamento alimentare.

In una fase iniziale sono stati esaminati dati storici relativi alla diffusione di tali patologie, con l'obiettivo di analizzarne l'evoluzione nel tempo e di individuare eventuali tendenze significative. Questa analisi esplorativa ha permesso di ottenere una visione complessiva della situazione attuale e delle dinamiche passate, supportata dall'utilizzo di grafici e strumenti di visualizzazione dei dati.

Successivamente sono state applicate tecniche di apprendimento non supervisionato per individuare gruppi di paesi con caratteristiche simili in termini di incidenza delle patologie mentali analizzate. Questo approccio ha consentito di identificare cluster omogenei di nazioni, facilitando la definizione di strategie di intervento differenziate e mirate.

Parallelamente sono state adottate tecniche di apprendimento supervisionato per la stima dei DALYs (Disability-Adjusted Life Years), un indicatore utilizzato per misurare l'impatto complessivo delle malattie sulla popolazione. I DALYs permettono di integrare sia gli anni di vita persi a causa di mortalità precoce sia gli anni vissuti con disabilità, fornendo una misura sintetica della gravità delle patologie considerate.

Infine, i dati e i risultati ottenuti sono stati organizzati attraverso un modello ontologico, con l'obiettivo di favorire la strutturazione, la condivisione e l'interoperabilità delle informazioni. L'integrazione con la Human Disease Ontology ha permesso di utilizzare definizioni e relazioni standardizzate già esistenti, rendendo i dati compatibili con altre fonti e strumenti che adottano la stessa ontologia e migliorando la validità complessiva dell'analisi.

ELENCO ARGOMENTI

Descrizione del Dominio e Dataset

- Origine dei dati
- Descrizione dataset usati
- Fonti e validità dei dati

Preprocessing dei Dati

- Pulizia dei dati
- Unione dei dataset
- Eliminazione e rinominazione colonne
- Standardizzazione e normalizzazione dei dati

Analisi Descrittiva

- Matrice di correlazione
- Andamento delle patologie dal 1990 al 2019
- Prevalenza e variabilità dei disturbi tra i diversi paesi nel mondo
- Visualizzazione grafica e interpretazione

Apprendimento supervisionato

- Obiettivi e preparazione dati
- Modelli predittivi usati:
    - Regressione lineare
    - Regressione polinomiale
    - Random Forest
    - Multi-Layer Perceptron (MLP)
- Ottimizzazione modelli e metriche di valutazione (RMSE, MAE, R^2)
- Risultati e analisi comparativa
- Previsioni future dei DALYs per disturbi mentali in Italia
- Visualizzazione e interpretazione delle previsioni

Apprendimento non supervisionato

- Modelli usati:
    - Agglomerative Clustering
    - KMeans Clustering
- Integrazione dei dati socioeconomici (GDP)
- Risultati e analisi cluster
    - Descrizione risultati ottenuti
    - Interpretazione risultati ottenuti
- Gruppi di intervento ottenuti

Ontologia

- Creazione dell'ontologia
- Integrazione con la Human Disease Ontology
- Mappatura dei disturbi mentali
- Creazione e serializzazione del grafo RDF

REQUISITI

Il codice in Python è stato progettato per poter essere eseguito direttamente sull'IDE, usando menù interattivi che danno la possibilità di scegliere tra diverse opzioni.
Le librerie usate sono:

- **Preprocessing**
  - pandas
  - sklearn

- **Visualizzazione grafici**
  - plotly
  - matplotlib
  - pandas
  - numpy
  - seaborn
  - geopandas
  - shapely

- **Apprendimento non supervisionato**
  - sklearn
  - KneeLocator

- **Apprendimento supervisionato**
  - Sklearn

- **Ontologie**
  - Rdflib

DESCRIZIONE DEL DOMINIO

Il dataset utilizzato per l'analisi è stato preso da Kaggle e contiene informazioni sulla prevalenza storica dei disturbi mentali e tassi di Disability-Adjusted Life Years (DALYs). Facendo una ricerca più approfondita è stato possibile verificare che i dati provengono dall'Organizzazione Mondiale della Sanità (OMS) e vengono usati per i rapporti sulla salute mentale, come evidenziato dalla scheda formativa che si può trovare su WHO. Oltre questo dataset è stato considerato anche un altro preso dalla World Bank Open Data: "World Development Indicators.csv", rinominato "SocialEconomicData.csv", il suo utilizzo verrà specificato nella sezione sull'apprendimento non supervisionato.

PREPROCESSING DEI DATI

Mental Health

Inizialmente i dati si dividevano in 7 dataset diversi:

1- mental-illnesses-prevalence

INDEX	ENTITY	CODE	YEAR	Schizophrenia disorders (share of population) - Sex: Both - Age: Age-standardized	Depressive disorders (share of population) - Sex: Both - Age: Age-standardized	Anxiety disorders (share of population) - Sex: Both - Age: Age-standardized	Bipolar disorders (share of population) - Sex: Both - Age: Age-standardized	Eating disorders (share of population) - Sex: Both - Age: Age-standardized
//	Stringa -Nome della Nazione,Paese o area	Stringa (DIM=3) identificativa delle nazioni	1990<=int<=2019	Double - Percentuale di popolazione. età standardizzata				

2- burden-disease-from-each-mental-illness

ENTITY	CODE	YEAR	DALYs - Sex: Both - Age: Age-standardized - Cause: Depressive disorders	DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Schizophrenia	DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Bipolar disorder	DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Eating disorders	DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Anxiety disorders
Stringa - Nome della Nazione, Paese o area	Stringa (DIM=3) identificativa delle nazioni	1990<=int<=2019	double - tasso di anni di vita sani persi per malattia ogni 100.000 persone				

Nel 2019, alcune regioni (con le migliori condizioni di salute) hanno registrato un tasso DALY inferiore a 20.000 ogni 100.000 persone, diversamente, nelle regioni più svantaggiate, il tasso si è dimostrato molto più elevato.
Per permettere un miglior confronto tra paesi nel tempo, questa metrica è stata standardizzata per età.

3-adult-population-covered-in-primary-data-on-the-prevalence-of-major-depression

Entity	Code	Year	Major depression
Stringa - Nome della Nazione,Paese o area	//	int - anno 2008	double - percentuale di copertura di ciascun paese per il quale sono stati raccolti dati su questa patologia es. se in un paese fossero stati raccolti dati sulla salute mentale di uomini e donne di tutte le fasce di età adulte, la copertura sarebbe pari al 100%.

4- adult-population-covered-in-primary-data-on-the-prevalence-of-mental-illnesses

ENTITY	CODE	YEAR	Major depression	Bipolar disorder	Eating disorders	Dysthymia	Schizophrenia	Anxiety disorders
Stringa - Regioni geografiche	Campi tutti vuoti tranne l'ultimo	Intero - anno 2008 in tutte le righe	Double - percentuale copertura di ciascun paese per il quale sono stati raccolti dati su questa patologia es. se in un paese fossero stati raccolti dati sulla salute mentale di uomini e donne di tutte le fasce di età adulte, la copertura sarebbe pari al 100%.					

5- anxiety-disorders-treatment-gap

ENTITY	CODE	YEAR	Potentially adequate treatment, conditional	Other treatments, conditional	Untreated, conditional
Stringa - Paese o categoria di paesi (es. high-income countries)	Stringa - 3 lettere che formano la sigla del paese. Alcuni campi sono vuoti.	Intero - anno (2002 - 2017)	Double - percentuale di persone con disturbi d'ansia che ha ricevuto un trattamento potenzialmente adeguato	Double - percentuale di persone con disturbi d'ansia che ha ricevuto altri tipi di trattamento	Double - percentuale di persone con disturbi d'ansia che non ha ricevuto un trattamento

6- depressive-symptoms-across-us-population

ENTITY	CODE	YEAR	Nearly every day	More than half the days	Several days	Not at all
Stringa - Sintomo (tutti diversi)	Colonna vuota	Intero - 2014 in tutte le righe	Double - percentuale di persone che, nelle due settimane precedenti, ha riscontrato il sintomo quasi tutti i giorni	Double - percentuale di persone che, nelle due settimane precedenti, ha riscontrato il sintomo per più della metà dei giorni	Double - percentuale di persone che, nelle due settimane precedenti, ha riscontrato il sintomo per molti giorni	Double - percentuale di persone che, nelle due settimane precedenti, non ha riscontrato il sintomo

Questi dati sono stati raccolti con un sondaggio sulla popolazione degli Stati Uniti nel 2014.

7-number-of-countries-with-primary-data-on-prevalence-of-mental-illnesses-in-theglobal-burden-of-disease-study

ENTITY	CODE	YEAR	Number of countries with primary data on prevalence of mental disorders
Stringa - Tipo di disturbo (tutti diversi)	Colonna vuota	Intero - 2019 in tutte le righe	Intero - da 2 a 172

Numero di paesi che hanno raccolto dati in un qualsiasi anno a partire dal 1980 sulla prevalenza di ciascun disturbo sulla popolazione.

Tra i 7 dataset disponibili, per rilevanza e semplicità nei calcoli, sono stati utilizzati:

1- mental-illnesses-prevalence
2- burden-disease-from-each-mental-illness

Questi dataset contengono I dati sulla prevalenza dei disturbi mentali e DALYs divisi per anno e area geografica, trattano i seguenti disturbi:

- Schizofrenia
- Depressione
- Ansia
- Disturbi alimentari
- Disturbi bipolari

Durante la fase di preprocessing:

- Sono state rinominate alcune colonne per semplicità;
- Sono state eliminate alcune colonne vuote;
- E' stato fatto il merge tra i due database, ottenendo un database unico "DisturbiMentali-DalysNazioniDelMondo", con lo scopo di ottenere un set di dati completo garantendo che non ci fossero informazioni ridondanti o duplicati.

Social Economic Data

Series Name	**Series Code**	**Country Name**	**Country Code**	**1990 [YR1990], 2000 [YR2000], 2014 [YR2014], 2015 [YR2015], 2016 [YR2016], 2017 [YR2017], 2018 [YR2018], 2019 [YR2019], 2020 [YR2020], 2021 [YR2021], 2022 [YR2022], 2023 [YR2023]**
Stringa - Nome della serie di dati un unico valore	Stringa - Codice della serie di dati un unico valore	Stringa - Nome della nazione a cui sono associati i valori	Stringa - Codice corrispondente alla nazione	Double - Valore del PIL per lo specifico anno

Durante la fase di preprocessing è stata rinominata la colonna “Country Code” in “Code” per coerenza con il dataset precedente e sono state eliminate le colonne vuote: “2022[YR2022]”, ”2023[YR2023]”.

Tutte le operazioni sono state eseguite usando Python e librerie come Pandas per manipolare i dati e filtrarli mantenendo informazioni pertinenti e utili per l’analisi.

FEATURES

Le features di questa fase di preprocessing sono:

- **DisturbiMentali-DalysNazioniDelMondo**

Nome colonna	Descrizione	Tipo
Entity	Rappresenta la nazione di riferimento.	String
Code	È il codice identificativo della nazione.	String
Year	Indica l'anno a cui si riferiscono i dati.	Int
Schizophrenia disorders	Rappresenta la prevalenza dei disturbi schizofrenici.	Double
Depressive disorders	Indica la prevalenza dei disturbi depressivi.	Double
Anxiety disorders	Indica la prevalenza dei disturbi d'ansia.	Double
Bipolar disorders	Rappresenta la prevalenza dei disturbi bipolari.	Double
Eating disorders	Indica la prevalenza dei disturbi alimentari.	Double
DALYs Cause: Depressive disorders	Rappresenta gli anni di vita persi per disabilità a causa di disturbi depressivi.	Double
DALYs Cause: Schizophrenia	Indica gli anni di vita persi per disabilità a	Double

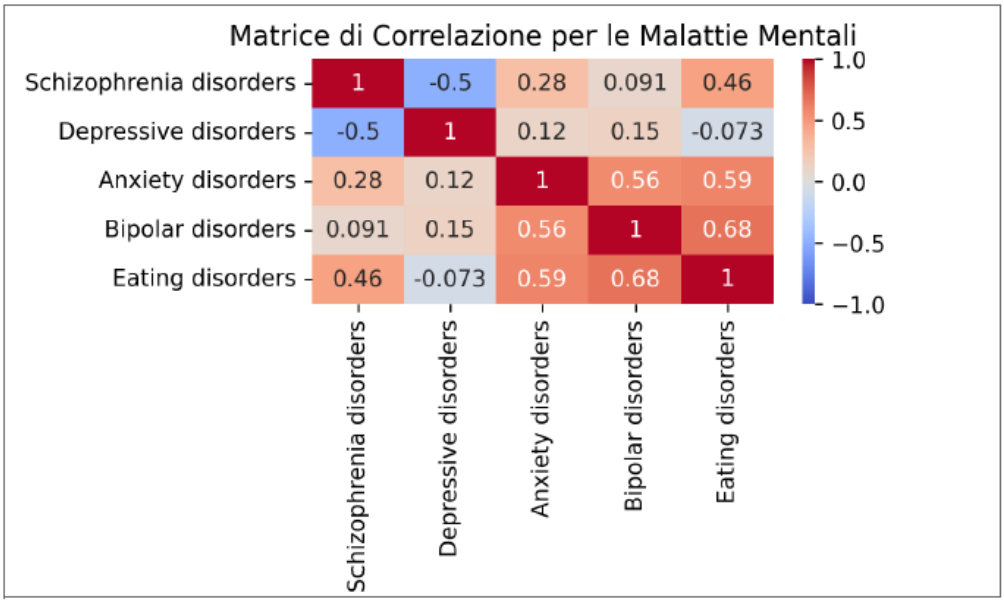
	causa di disturbi schizofrenici.	
DALYs Cause: Bipolar disorder	Rappresenta gli anni di vita persi per disabilità a causa di disturbi bipolari.	Double
DALYs Cause: Eating disorders	Indica gli anni di vita persi per disabilità a causa di disturbi alimentari.	Double
DALYs Cause: Anxiety disorders	Rappresenta gli anni di vita persi per disabilità a causa di disturbi d'ansia.	Double

- **SocialEconomicData**

Nome colonna	**Descrizione**	**Tipo**
Series Name	Nome della serie di dati	Stringa
Series Code	Codice della serie di dati	Stringa
Country Name	Nome della nazione a cui sono associati i valori	Stringa
Code	Codice corrispondente alla nazione	Stringa
1990 [YR1990]	Valore del PIL per l'anno 1990	Double
2000 [YR2000]	Valore del PIL per l'anno 2000	Double
2014 [YR2014]	Valore del PIL per l'anno 2014	Double
2015 [YR2015]	Valore del PIL per l'anno 2015	Double

2016 [YR2016]	Valore del PIL per l'anno 2016	Double
2017 [YR2017]	Valore del PIL per l'anno 2017	Double
2018 [YR2018]	Valore del PIL per l'anno 2018	Double
2019 [YR2019]	Valore del PIL per l'anno 2019	Double
2020 [YR2020]	Valore del PIL per l'anno 2020	Double
2021 [YR2021]	Valore del PIL per l'anno 2021	Double

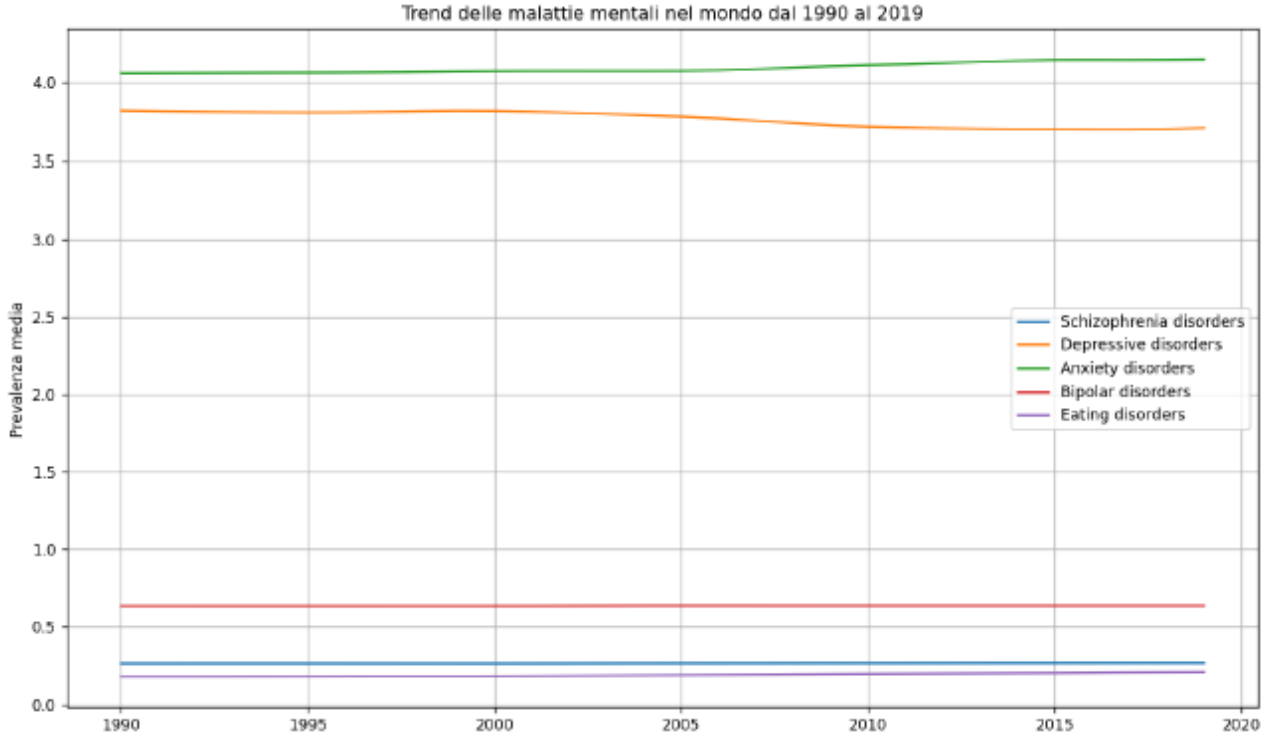
ANALISI DESCRITTIVA DATASET



- **Disturbi depressivi:**
  - Si osservano correlazioni positive con i disturbi bipolari (0.147) e d'ansia (0.115);
  - È presente una correlazione negativa con i disturbi schizofrenici (-0.499).
- **Disturbi schizofrenici:**
  - Presentano correlazioni positive con i disturbi d'ansia (0.28) e alimentari (0.46);
  - È rilevata una correlazione negativa con i disturbi depressivi (-0.499).
- **Disturbi d'ansia:**
  - Mostrano correlazioni positive con i disturbi alimentari (0.59) e bipolari (0.56);
  - Sono presenti correlazioni negative con i disturbi schizofrenici (-0.26) e depressivi (-0.115).

- **Disturbi alimentari:**
  - Si osservano correlazioni positive con i disturbi d'ansia (0.59) e bipolari (0.67);
  - È presente una correlazione negativa con i disturbi depressivi (-0.073).
- **Disturbi bipolari:**
  - Presentano correlazioni positive con i disturbi d'ansia (0.56) e alimentari (0.67);
  - Non emergono correlazioni negative rilevanti

ANDAMENTO MEDIO PATOLOGIE DAL 1990 AL 2019



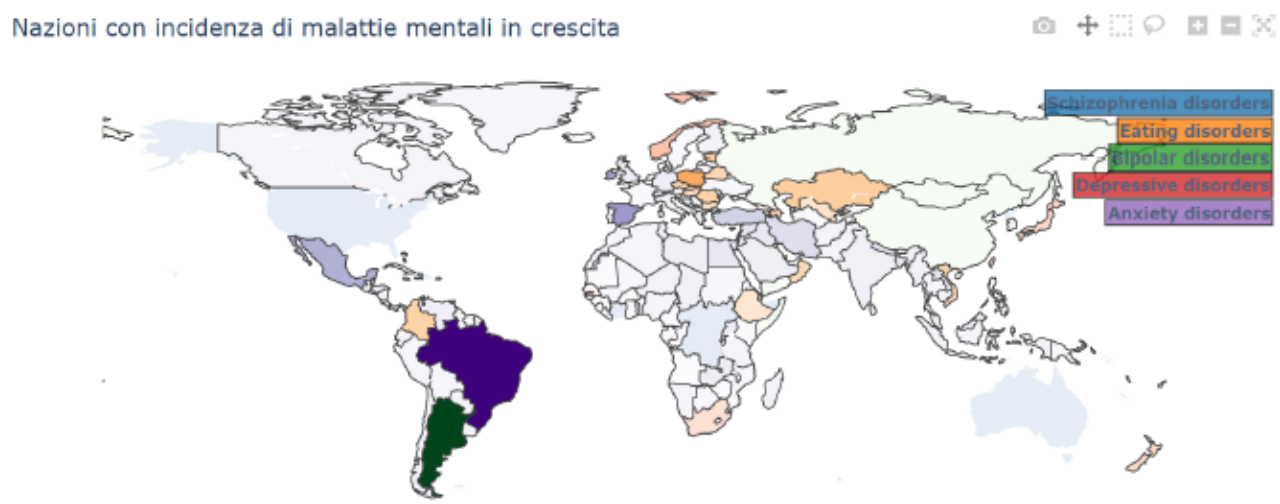
Ho creato questo grafico per analizzare l'andamento della prevalenza media dei principali disturbi mentali nel periodo 1990–2019.
Dal grafico emergono le seguenti osservazioni:

- **Disturbi d'ansia:** presentano la prevalenza media più elevata rispetto agli altri disturbi considerati e mostrano un lieve incremento nel tempo;
- **Disturbi depressivi:** mostrano una prevalenza media relativamente stabile nel periodo analizzato, inferiore rispetto ai disturbi d'ansia;
- **Disturbi bipolari:** presentano una prevalenza media bassa e nettamente inferiore rispetto ai disturbi d'ansia e depressivi;
- **Disturbi schizofrenici:** mostrano una prevalenza media stabile e rappresentano il disturbo con la prevalenza più bassa;
- **Disturbi alimentari:** presentano una prevalenza media contenuta e una tendenza sostanzialmente costante nel tempo.

Nel complesso, dal grafico si osserva che i disturbi d'ansia e depressivi risultano essere i più diffusi a livello globale nel periodo considerato, mentre i disturbi schizofrenici e alimentari mostrano valori medi inferiori.

PREVALENZA E VARIABILITA' DEI DISTURBI TRA I DIVERSI PAESI

La prevalenza media dei **disturbi schizofrenici** risulta pari a circa **0,27%**, con una deviazione standard contenuta. Questo indica una diffusione relativamente bassa e una variabilità limitata tra i diversi paesi.
La prevalenza media dei **disturbi depressivi** è pari a circa **3,76%**, con una deviazione standard elevata. Questo dato suggerisce una marcata eterogeneità tra le nazioni, evidenziando differenze significative nella diffusione del disturbo a livello globale.
I **disturbi d'ansia** presentano la prevalenza media più elevata, pari a circa **4,09%**, anch'essa caratterizzata da un'elevata deviazione standard. Anche in questo caso, la variabilità osservata indica una distribuzione non uniforme tra i diversi paesi.
Per quanto riguarda i **disturbi bipolari**, il dataset evidenzia una sostanziale stabilità nel tempo, mentre i **disturbi alimentari** mostrano una tendenza all'aumento in diverse aree geografiche.
Per identificare le nazioni che presentano un incremento della prevalenza dei disturbi alimentari, è stata analizzata la tendenza temporale della prevalenza per ciascun paese mediante un modello di regressione lineare. In particolare, è stata calcolata la **pendenza della retta di regressione**, che rappresenta il tasso di variazione annuale del disturbo. È stata considerata esclusivamente la pendenza, in quanto indicativa dell'andamento nel tempo: valori positivi indicano un aumento della prevalenza, mentre valori negativi suggeriscono una diminuzione. Successivamente, sono stati selezionati solo i paesi caratterizzati da una pendenza positiva.
Infine, i risultati sono stati rappresentati tramite una mappa geografica che evidenzia le nazioni con una tendenza crescente nella prevalenza dei disturbi mentali, consentendo una visualizzazione immediata delle aree maggiormente interessate dal fenomeno.



APPRENDIMENTO SUPERVISIONATO

Per la previsione del carico futuro dei disturbi mentali (depressivi, schizofrenici, bipolari, alimentari e d'ansia) in Italia sono state applicate tecniche di apprendimento supervisionato.

Preparazione dei dati:
Il dataset è stato inizialmente filtrato per considerare esclusivamente i dati relativi all'Italia. Successivamente sono state selezionate le variabili di interesse, includendo l'anno e i valori dei DALYs per ciascun disturbo mentale. È stato infine definito un intervallo temporale futuro (2020–2030) per stimare l'evoluzione dei DALYs nel prossimo decennio.

Modelli utilizzati

- **Regressione lineare**
  Il modello assume una relazione lineare tra la variabile indipendente *Year* e la variabile dipendente *DALYs*.
  La valutazione delle prestazioni è stata effettuata tramite cross-validation a 5 fold, suddividendo il dataset in cinque parti, utilizzandone quattro per l'addestramento e una per la validazione, ripetendo il processo cinque volte.
  La metrica di valutazione utilizzata è l'RMSE (Root Mean Squared Error).
- **Regressione polinomiale (grado 2)**
  Questo modello estende la regressione lineare introducendo termini polinomiali di secondo grado, consentendo di catturare relazioni non lineari tra le variabili.
  Anche in questo caso è stata applicata la cross-validation a 5 fold, utilizzando l'RMSE come metrica di confronto con il modello lineare, al fine di valutare l'effettivo miglioramento delle prestazioni predittive.
- **Random Forest**
  Modello di tipo ensemble basato su alberi decisionali, in grado di modellare relazioni complesse e non lineari nei dati.
  Per ottimizzare le prestazioni del modello è stata applicata una procedura di Randomized Search sui principali iperparametri, tra cui:
  - numero di alberi (*n_estimators*),
  - profondità massima degli alberi (*max_depth*),
  - numero minimo di campioni per la suddivisione di un nodo (*min_samples_split*),
  - numero minimo di campioni per un nodo foglia (*min_samples_leaf*).

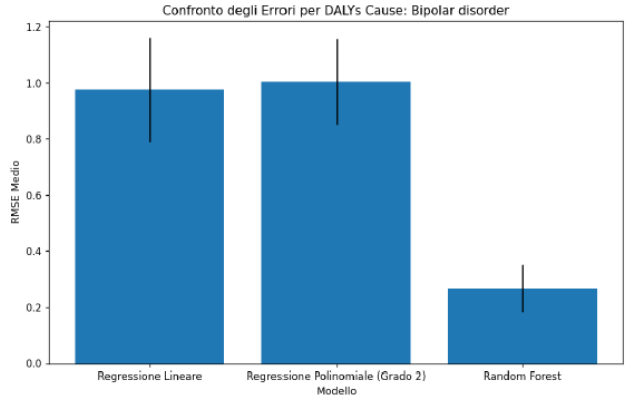
Anche per questo modello è stata utilizzata la cross-validation a 5 fold con RMSE come metrica di valutazione.
L'errore quadratico medio (RMSE) è stato scelto come metrica principale in quanto penalizza maggiormente gli errori di previsione più elevati, risultando particolarmente adatto per valutare l'accuratezza dei modelli nel contesto dei DALYs.

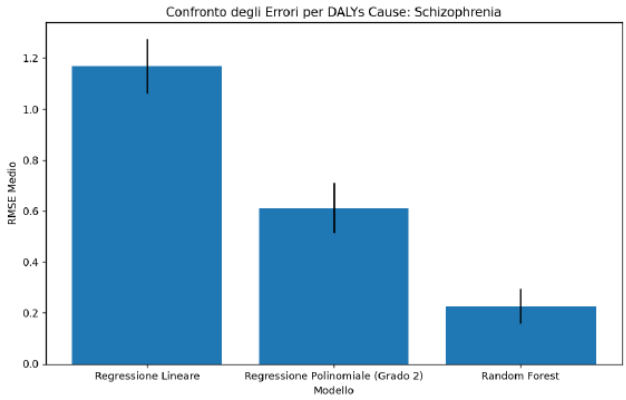
ANALISI DEI RISULTATI

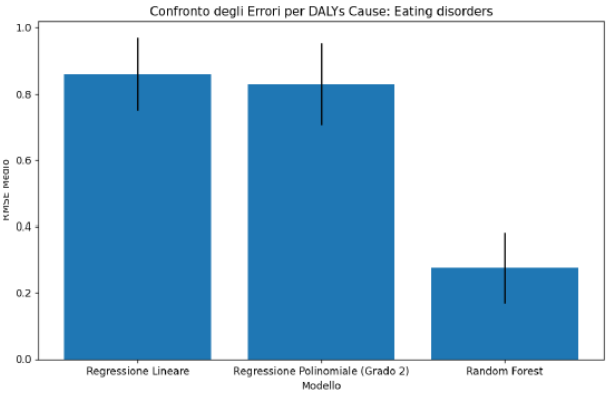
Risultati del Confronto dei Modelli

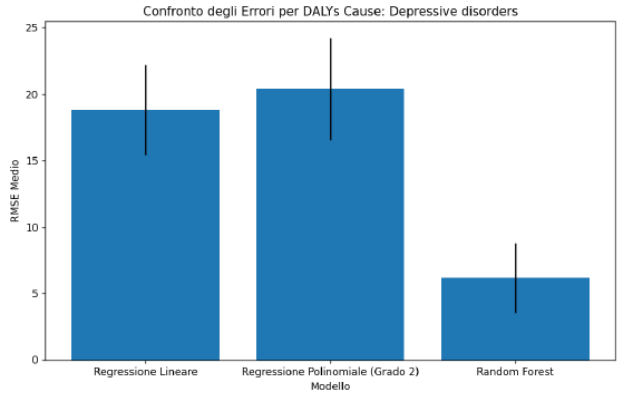
	Modello	Disturbo	RMSE Medio	Deviazione Std RMSE	Migliori Parametri
0	Regressione Lineare	DALYs Cause: Depressive disorders	18.802480	3.400771	None
1	Regressione Polinomiale (Grado 2)	DALYs Cause: Depressive disorders	20.392532	3.840496	None
2	Random Forest	DALYs Cause: Depressive disorders	6.186234	2.618246	{'n_estimators': 100, 'min_samples_split': 5, ...
3	Regressione Lineare	DALYs Cause: Schizophrenia	1.169332	0.106849	None
4	Regressione Polinomiale (Grado 2)	DALYs Cause: Schizophrenia	0.612539	0.099318	None
5	Random Forest	DALYs Cause: Schizophrenia	0.227337	0.069524	{'n_estimators': 200, 'min_samples_split': 2, ...
6	Regressione Lineare	DALYs Cause: Bipolar disorder	0.976395	0.186807	None
7	Regressione Polinomiale (Grado 2)	DALYs Cause: Bipolar disorder	1.004136	0.153276	None
8	Random Forest	DALYs Cause: Bipolar disorder	0.267820	0.085154	{'n_estimators': 200, 'min_samples_split': 5, ...
9	Regressione Lineare	DALYs Cause: Eating disorders	0.859754	0.109899	None
10	Regressione Polinomiale (Grado 2)	DALYs Cause: Eating disorders	0.829097	0.123590	None
11	Random Forest	DALYs Cause: Eating disorders	0.275430	0.106233	{'n_estimators': 200, 'min_samples_split': 2, ...
12	Regressione Lineare	DALYs Cause: Anxiety disorders	29.588484	6.297473	None
13	Regressione Polinomiale (Grado 2)	DALYs Cause: Anxiety disorders	17.446128	2.509318	None
14	Random Forest	DALYs Cause: Anxiety disorders	6.005097	2.942714	{'n_estimators': 200, 'min_samples_split': 2, ...

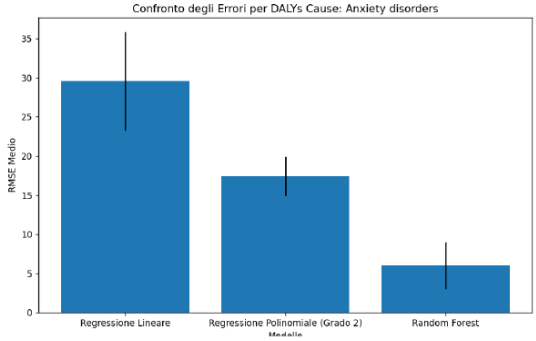
Dai risultati ottenuti emerge che i modelli **Random Forest** presentano valori di RMSE mediamente inferiori rispetto ai modelli di regressione lineare e polinomiale, indicando prestazioni predittive superiori.
I grafici a barre riportano il confronto dell'RMSE medio per ciascun disturbo mentale e per ciascun modello considerato. Valori più bassi di RMSE indicano una maggiore accuratezza del modello, evidenziando come l'approccio Random Forest risulti complessivamente il più efficace nella maggior parte dei casi analizzati.











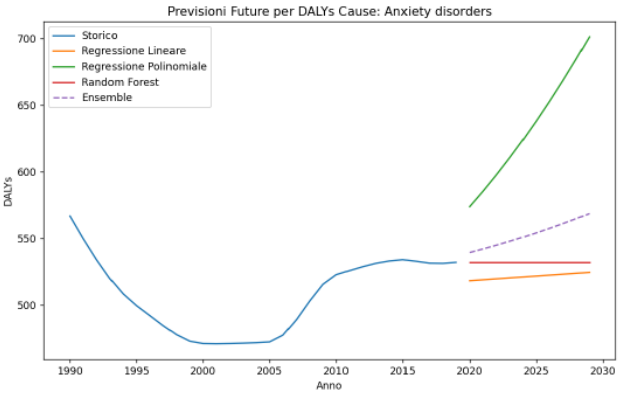
APPROCCIO ENSEMBLE

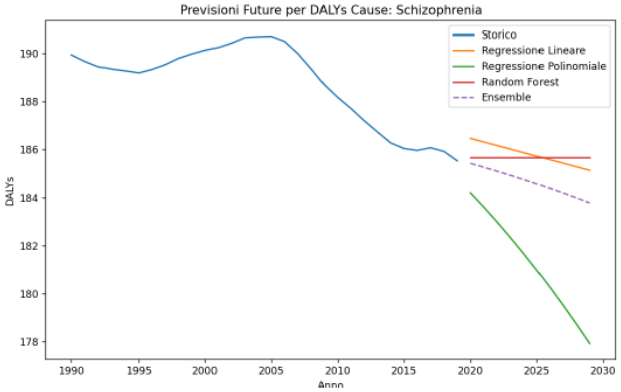
È stato implementato un approccio **ensemble** al fine di combinare le previsioni ottenute dai tre modelli supervisionati considerati (regressione lineare, regressione polinomiale e Random Forest).
La previsione finale dell'ensemble è stata calcolata come **media pesata** delle singole previsioni, assegnando a ciascun modello un peso inversamente proporzionale al rispettivo valore di **RMSE**. In questo modo, i modelli con prestazioni migliori contribuiscono maggiormente alla stima finale.
L'obiettivo dell'approccio ensemble è quello di ottenere una previsione più stabile ed equilibrata, riducendo la varianza e sfruttando i punti di forza dei diversi modelli.
Dai risultati ottenuti emerge che il modello ensemble fornisce prestazioni competitive per alcuni disturbi mentali, tuttavia **non supera sistematicamente le prestazioni del modello Random Forest**, che risulta generalmente il più accurato in termini di RMSE.

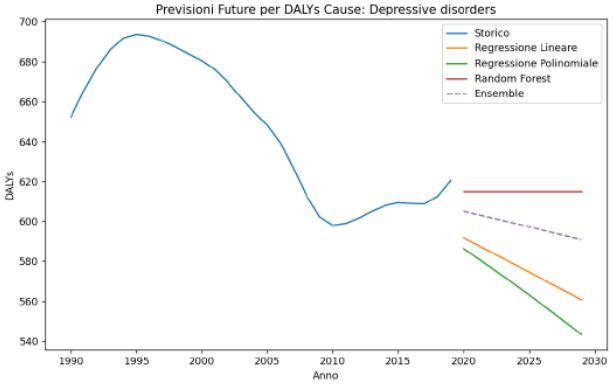
Risultati del Confronto dei Modelli Ensemble

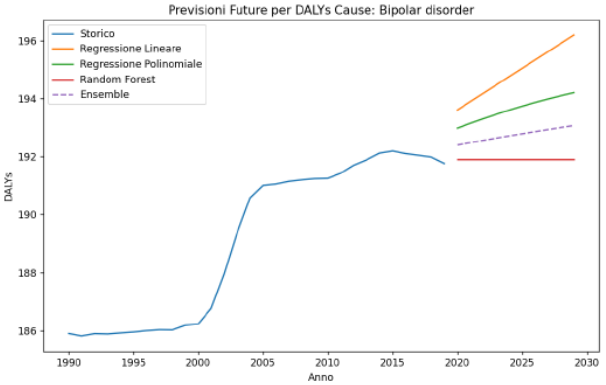
	Disturbo	RMSE Medio Ensemble	Deviazione Std RMSE Ensemble
0	DALYs Cause: Depressive disorders	17.498306	3.013694
1	DALYs Cause: Schizophrenia	1.055574	0.213315
2	DALYs Cause: Bipolar disorder	1.061619	0.215637
3	DALYs Cause: Eating disorders	0.762419	0.179685
4	DALYs Cause: Anxiety disorders	27.831985	8.658387

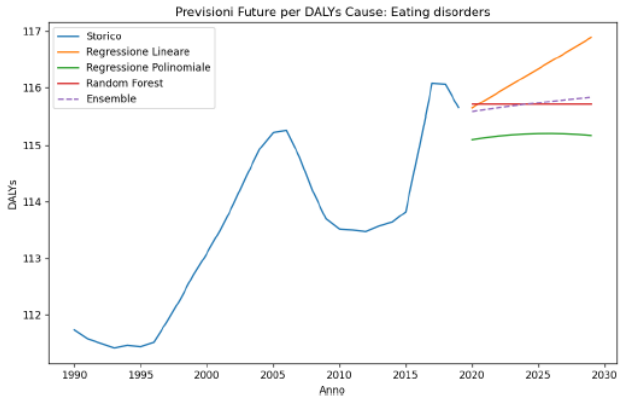
In particolare, per i **disturbi d'ansia**, il modello ensemble presenta un valore di RMSE più elevato rispetto al Random Forest, indicando una minore precisione predittiva. Questo risultato suggerisce che, in presenza di relazioni non lineari complesse, il modello Random Forest riesce a catturare meglio la dinamica dei dati rispetto alla combinazione dei modelli lineari e polinomiali.
I grafici delle previsioni future mostrano il confronto tra l'andamento storico dei DALYs e le stime prodotte dai singoli modelli e dall'ensemble per il periodo 2020–2030, evidenziando differenze significative nelle traiettorie previste, in particolare per i disturbi caratterizzati da maggiore variabilità nel tempo.











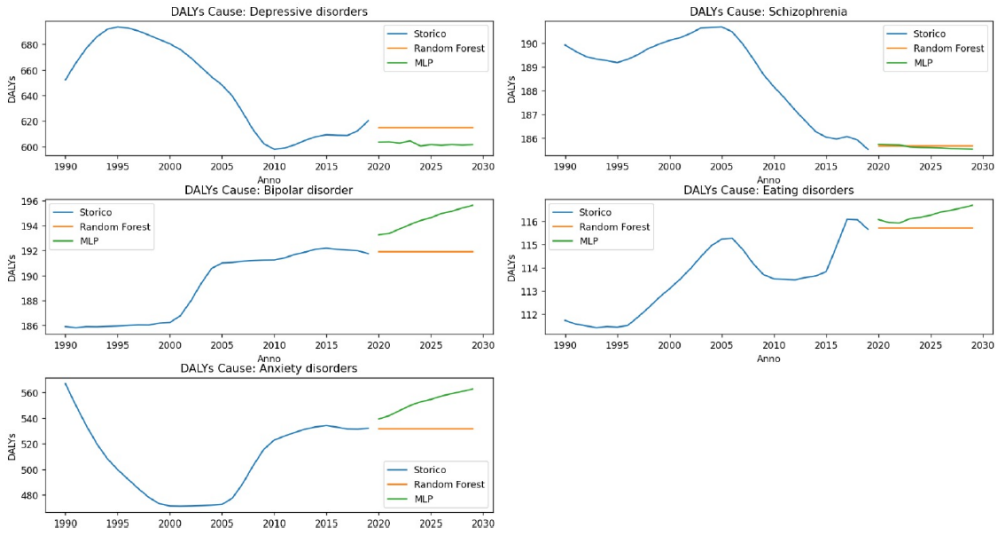
RANDOM FOREST VS MLP(Multi-Layer Perceptron)

Confrontando i modelli Random Forest e MLP per la previsione dei DALYs associati ai vari disturbi mentali in Italia, emergono alcuni elementi importanti:

- Per tutti i disturbi considerati il modello Random Forest ha mostrato un errore quadratico medio inferiore rispetto all'MLP, questo indica maggior precisione nelle previsioni;
- L'errore assoluto medio (MAE) è stato inferiore per Random Forest rispetto all'MLP;
- Il coefficiente di determinazione (R^2) ha assunto valori negativi in entrambi i modelli, indicando che le prestazioni predittive risultano inferiori a quelle di un modello che predice semplicemente la media dei dati. Tuttavia, il Random Forest presenta valori di R^2 sistematicamente migliori rispetto all'MLP, seppur negativi.

	Disorder	Random Forest RMSE	MLP RMSE	Random Forest MAE	MLP MAE	Random Forest R2	MLP R2	Random Forest MAPE	MLP MAPE
0	DALYs Cause: Depressive disorders	10.131990	8.693553	8.973132	7.189237	-1.557115	-0.882589	0.014852	0.011773
1	DALYs Cause: Schizophrenia	1.200769	1.196534	0.908052	0.934635	-1.150235	-1.135094	0.004849	0.004992
2	DALYs Cause: Bipolar disorder	0.304171	2.695697	0.241748	2.621019	-0.049443	-81.426365	0.001261	0.013660
3	DALYs Cause: Eating disorders	1.676882	2.015996	1.432395	1.829900	-1.460087	-2.555698	0.012589	0.016062
4	DALYs Cause: Anxiety disorders	3.684631	22.626503	2.412888	21.989455	-0.207538	-44.535250	0.004580	0.041424

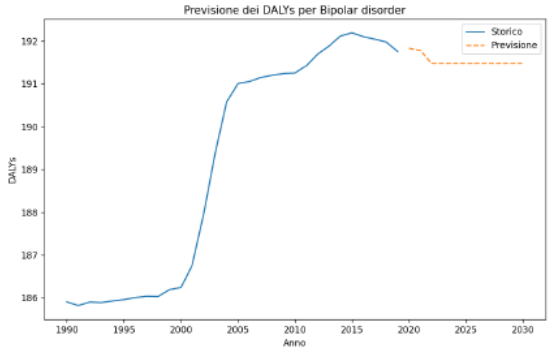
PS C:\Users\letha\Desktop\ICON descrizioni>

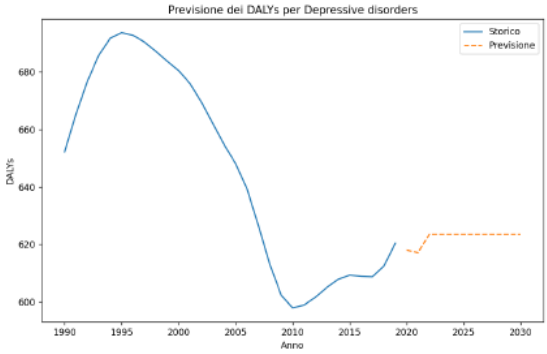


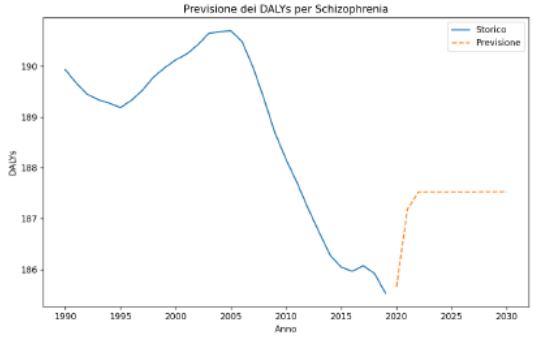
OTTIMIZZAZIONE RANDOM FOREST

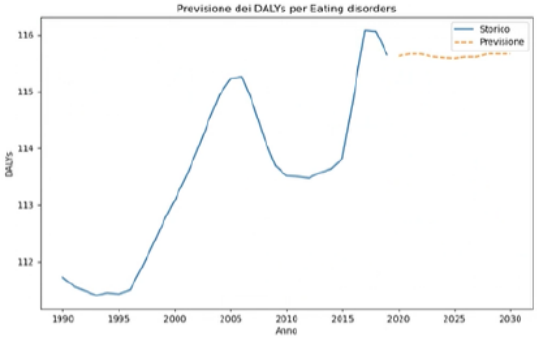
Per migliorare il modello e le sue performance, è stata usata una nuova strategia di modellazione del Random Forest.
Inizialmente è stata eseguita una selezione delle feature, prendendo in considerazione tutte le metriche disponibili nel dataset e aggiungendo l'anno come variabile di input, questo passaggio è molto importante in quanto aggiungere delle variabili può fornire informazioni utili al modello.
Per valutare il modello è stata usata una tecnica di cross-validation con K-Fold.
Successivamente, è stata applicata una standardizzazione delle variabili tramite *StandardScaler* per garantire coerenza con il processo di validazione e mantenere uniformità con gli altri modelli considerati, pur non essendo strettamente necessaria per il Random Forest.
Il modello è stato valutato usando le metriche: R^2, MSE, RMSE, MAE.
Per fare ciò, sono stati utilizzati i dati storici disponibili per l'Italia come input e proiettando le stime sui periodi futuri, senza reinserire nel training dati predetti.
L'ottimizzazione del modello ha portato a un miglioramento significativo delle prestazioni predittive, come evidenziato dai valori elevati di R^2 e dalla riduzione degli errori RMSE e MAE, confermando l'efficacia del Random Forest per la previsione dei DALYs.

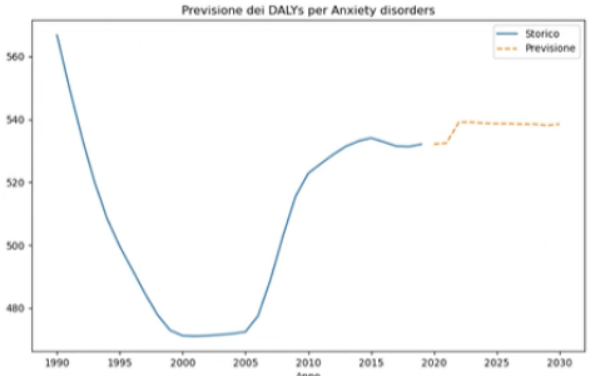
```
Valori di valutazione per DALYs Cause: Depressive disorders:
R²: 0.994206265100252
MSE: 6.310292467272926
RMSE: 2.512029551433049
MAE: 2.255182633333883
Cross-Validation RMSE Mean: 4.391850077435272
Cross-Validation RMSE Std: 0.910561028535973

Valori di valutazione per DALYs Cause: Schizophrenia:
R²: 0.9667089351808889
MSE: 0.10559168251257384
RMSE: 0.3249487382843236
MAE: 0.25422249999999263
Cross-Validation RMSE Mean: 0.28320515351668085
Cross-Validation RMSE Std: 0.12180938351103354

Valori di valutazione per DALYs Cause: Bipolar disorder:
R²: 0.9872959887032145
MSE: 0.08467681558357053
RMSE: 0.29099281019222883
MAE: 0.20096925000004262
Cross-Validation RMSE Mean: 0.45512259143537814
Cross-Validation RMSE Std: 0.25227797463123225

Valori di valutazione per DALYs Cause: Eating disorders:
R²: 0.9464306217444229
MSE: 0.09975614786423082
RMSE: 0.3158419665975863
MAE: 0.26044140500004903
Cross-Validation RMSE Mean: 0.29890463372206905
Cross-Validation RMSE Std: 0.11067009374911922

Valori di valutazione per DALYs Cause: Anxiety disorders:
R²: 0.995930089453654
MSE: 2.6951149674878585
RMSE: 1.6416805314944374
MAE: 1.1302157333334435
Cross-Validation RMSE Mean: 5.337204962987887 5
Cross-Validation RMSE Std: 2.925589309258347
```











CONCLUSIONI

Le previsioni sui disturbi depressivi indicano un lieve incremento dei DALYs nei prossimi anni, potenzialmente associato all'invecchiamento della popolazione e a cambiamenti nello stile di vita.

Per la schizofrenia emerge un possibile picco dei DALYs, suggerendo un aumento dell'impatto complessivo sulla popolazione.

I disturbi alimentari e bipolari mostrano una crescita contenuta, mentre i disturbi d'ansia evidenziano una tendenza più marcata all'aumento, plausibilmente legata a fattori di stress e instabilità socio-economica.

APPRENDIMENTO NON SUPERVISIONATO

L'obiettivo è identificare cluster distinti di paesi con profili di salute mentale simili.
Sono stati usati due metodi di clustering: **Agglomerative clustering e KMeans Clustering,** per confrontarli e determinare quale metodo fornisce una migliore separazione dei cluster.
Per essere sicura che il confronto sia equo tra le nazioni, i dati relativi i disturbi mentali sono stati normalizzati usando la tecnica **StandardScaler.**
Durante l'analisi, è stata necessaria l'integrazione dei dati socioeconomici del dataset dalla World Bank, World Development Indicators.csv. In particolare, il GDP, Prodotto Interno Lordo (Gross Domestic Product in inglese), una misura del valore monetario totale di tutti i beni e servizi prodotti all'interno di un paese durante un periodo di tempo specifico.

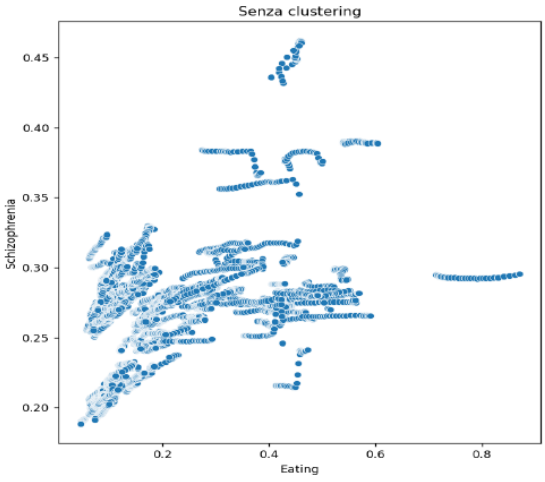
CLUSTERING AGGLOMERATIVO

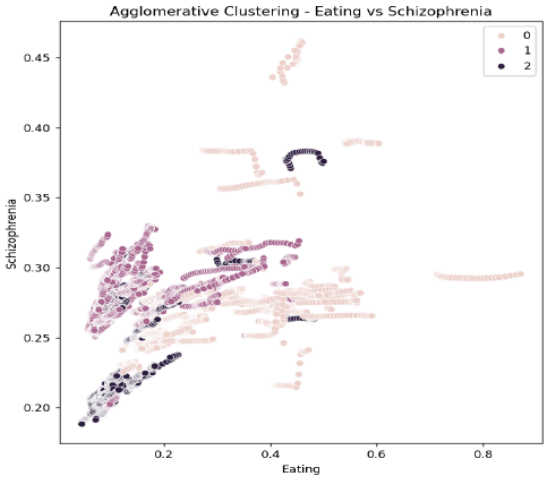
È una tecnica di clustering gerarchico che costruisce un albero chiamato dendrogramma unendo iterativamente i punti dati simili.
Si parte con ogni punto dato come un cluster separato e si procede unendo i cluster simili, fino ad ottenere uno unico.

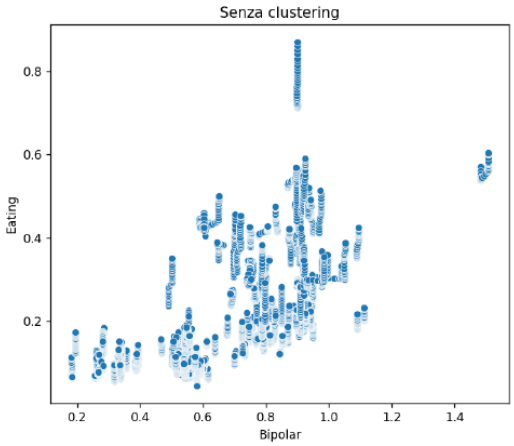
Il clustering agglomerativo è stato eseguito utilizzando il metodo di collegamento ward (**linkage="ward"**) che minimizza la varianza dei cluster durante la fusione, i risultati sono:

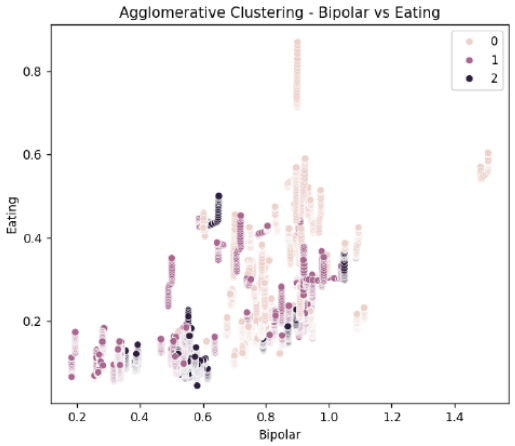
- **Cluster 0:** include paesi occidentali e dell'America Latina;
- **Cluster 1:** include molte nazioni dell'Asia, est Europa e regioni Pacifico;
- **Cluster 2:** include paesi africani.

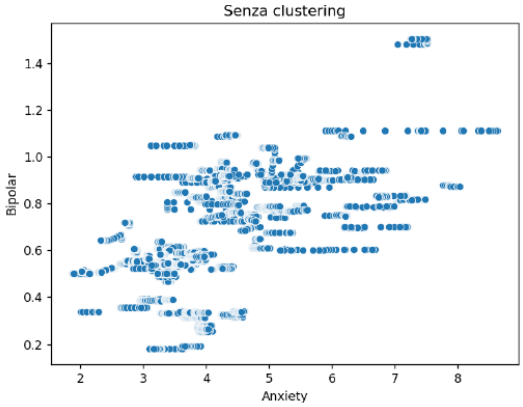
Questo modello ha prodotto un coefficiente di silhouette di 0.388 che serve a misurare la qualità della separazione dei cluster.

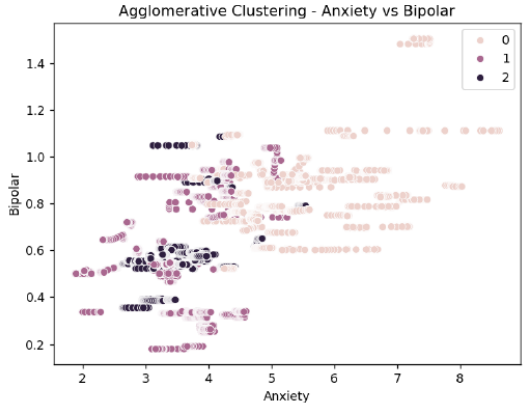


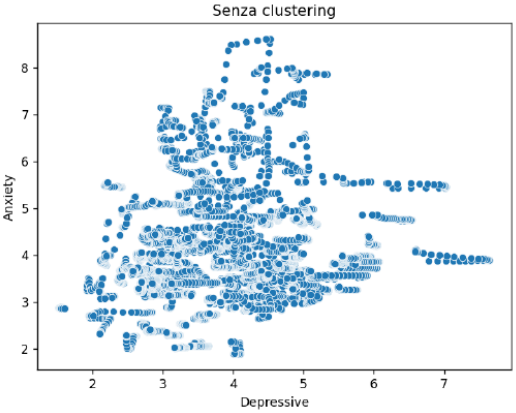


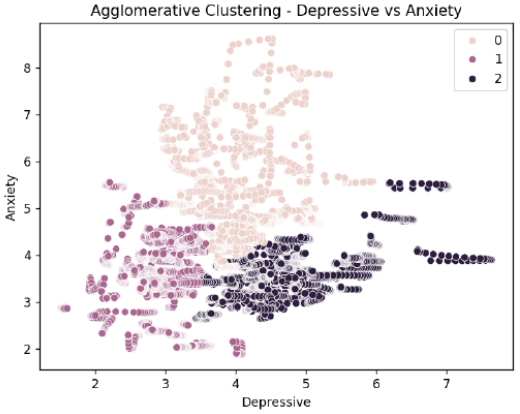

Senza clustering
Eating
Bipolar

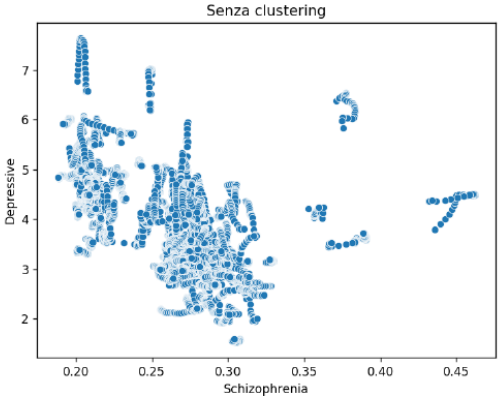

Agglomerative Clustering - Bipolar vs Eating
0
1
2
Eating
Bipolar

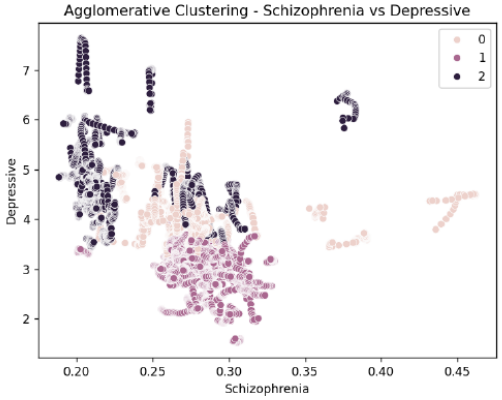

Senza clustering
Bipolar
Anxiety

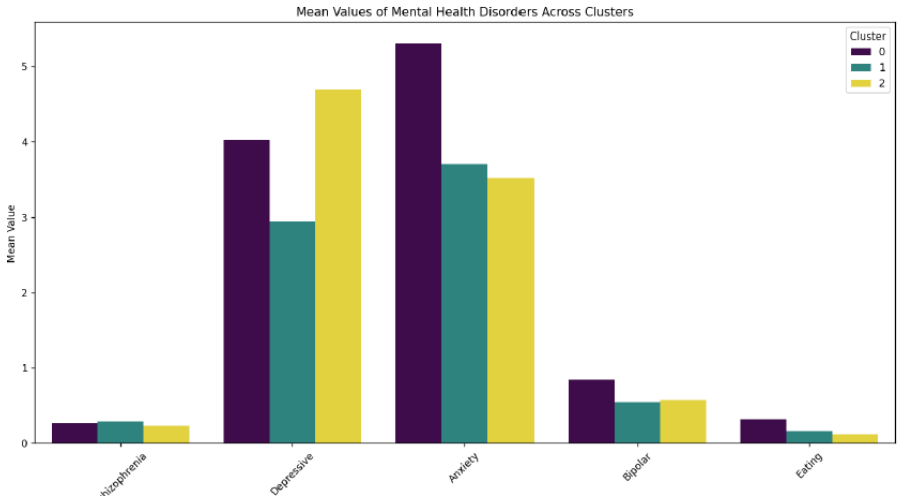

Agglomerative Clustering - Anxiety vs Bipolar
0
1
2
Bipolar
Anxiety


Senza clustering
Anxiety
Depressive


Agglomerative Clustering - Depressive vs Anxiety
0
1
2
Anxiety
Depressive

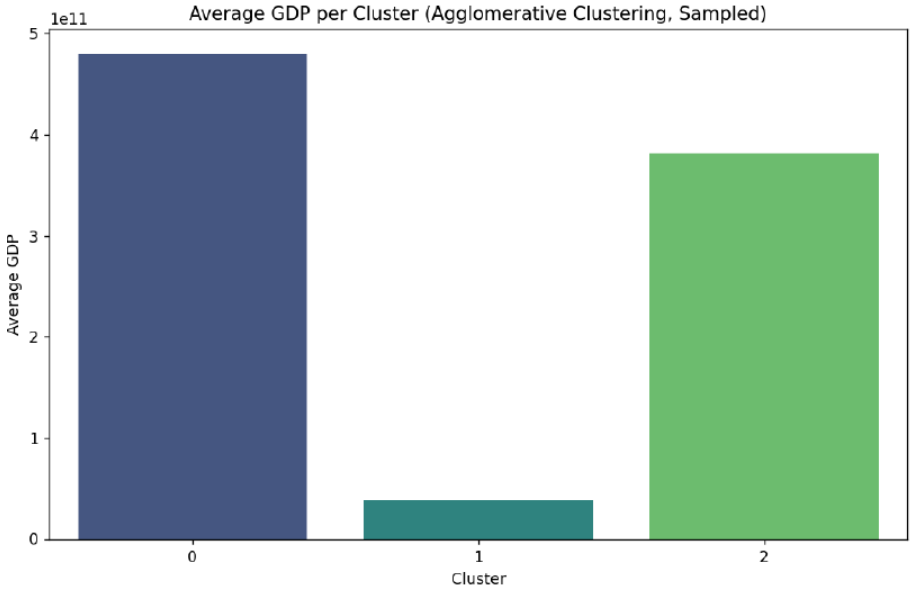

Senza clustering
Depressive
Schizophrenia


Agglomerative Clustering - Schizophrenia vs Depressive
0
1
2
Depressive
Schizophrenia



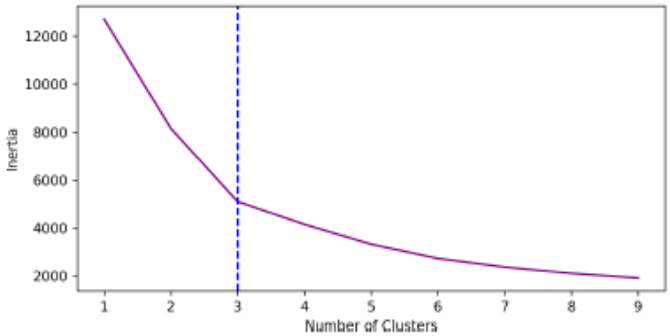
Sono stati considerati altri fattori nell'analisi, come il GDP, questo ha portato a nuove conclusioni:

- Le nazioni con GDP alto tendono ad avere prevalenza maggiore di disturbi mentali, probabilmente a causa di una migliore diagnosi e accesso ai servizi,
- Le nazioni con un GDP basso, anche se hanno una minore prevalenza registrata, sono soggette a mancanze di risorse per diagnosi e trattamenti.

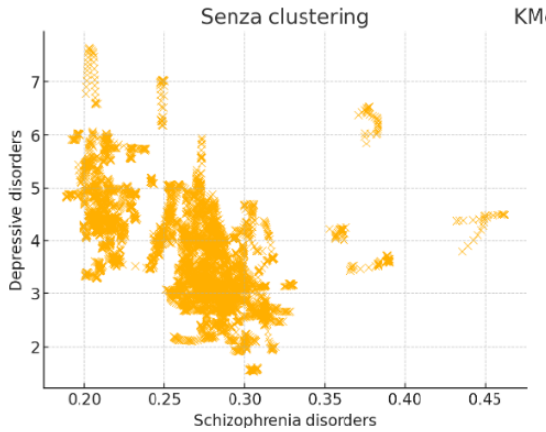


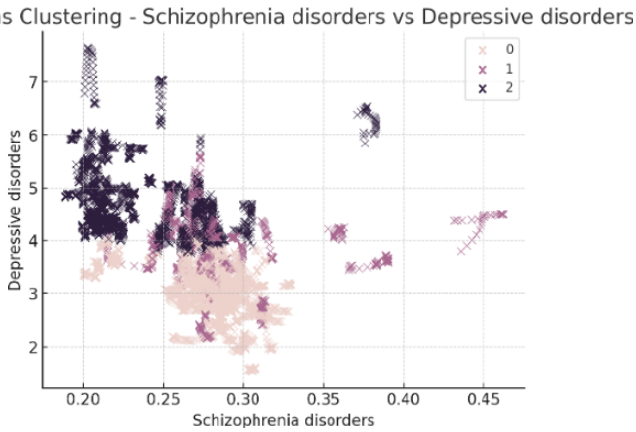
KMeans CLUSTERING

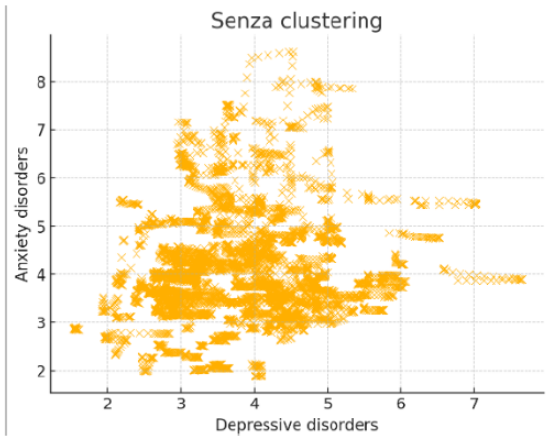
È stato usato il metodo dell'elbow (gomito) per determinare il numero ottimale di cluster in un'analisi K-means. Grazie a questo metodo ho potuto identificare il punto in cui l'aggiunta di altri cluster non porta a un miglioramento nelle qualità del clustering. Analizzando la somma delle distanze quadrate interne ai cluster, ho riscontrato che il numero ottimale è 3.

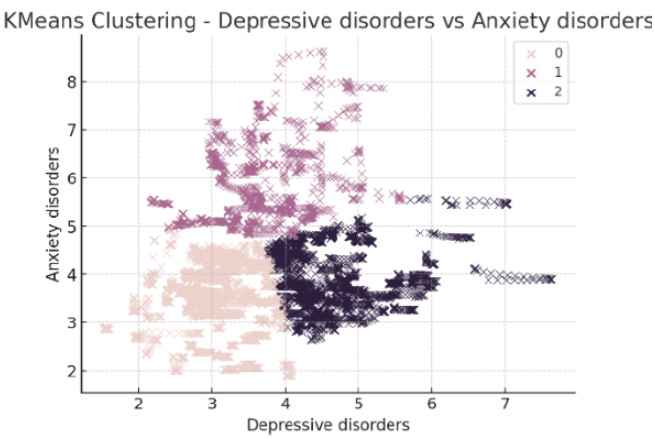


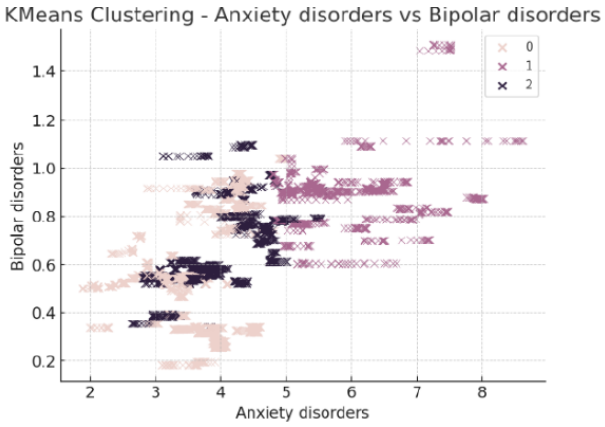
Il risultato del coefficiente di silhouette per il clustering KMeans è leggermente superiore, pari a 0.400, questo significa che il metodo KMeans fornisce una separazione dei clustering leggermente migliore rispetto al metodo agglomerativo.

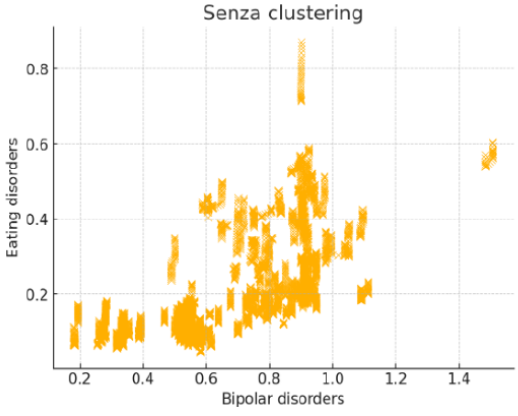


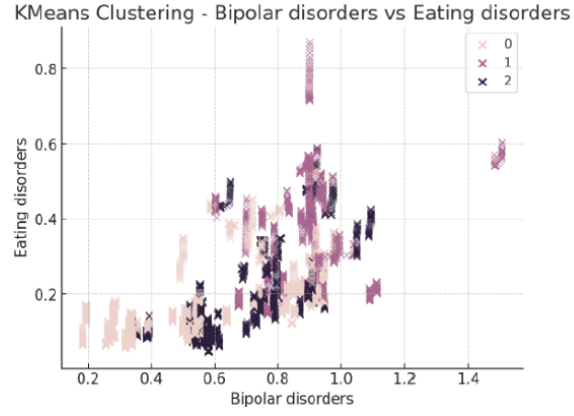


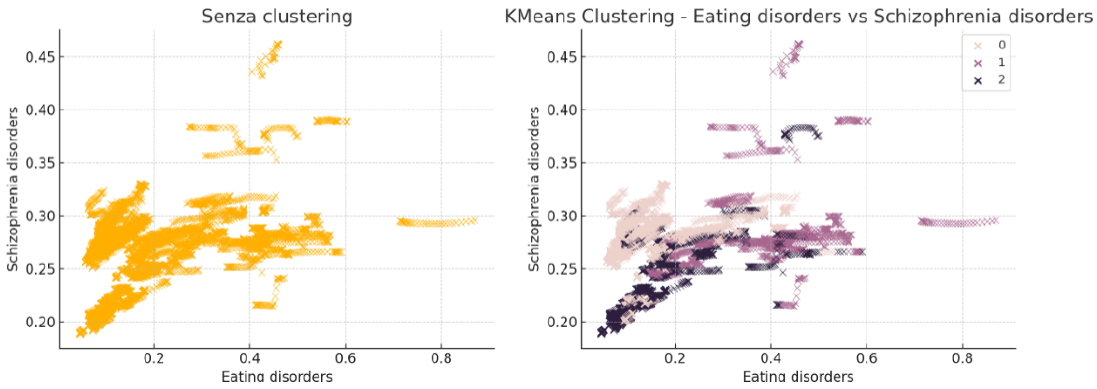


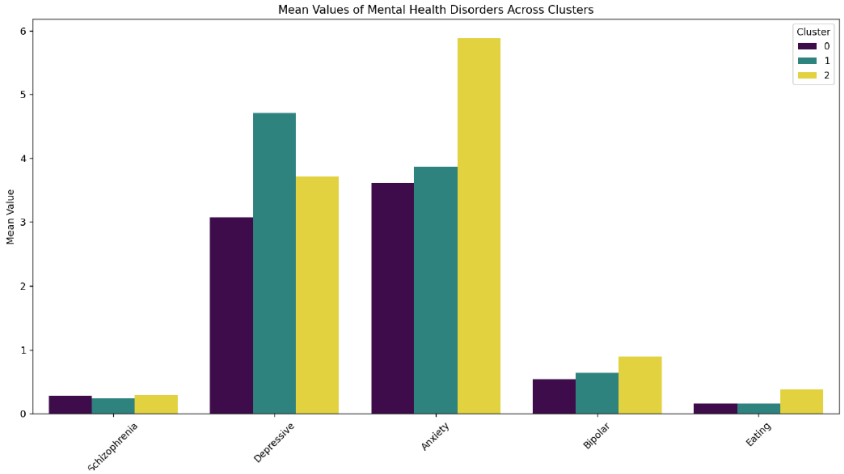






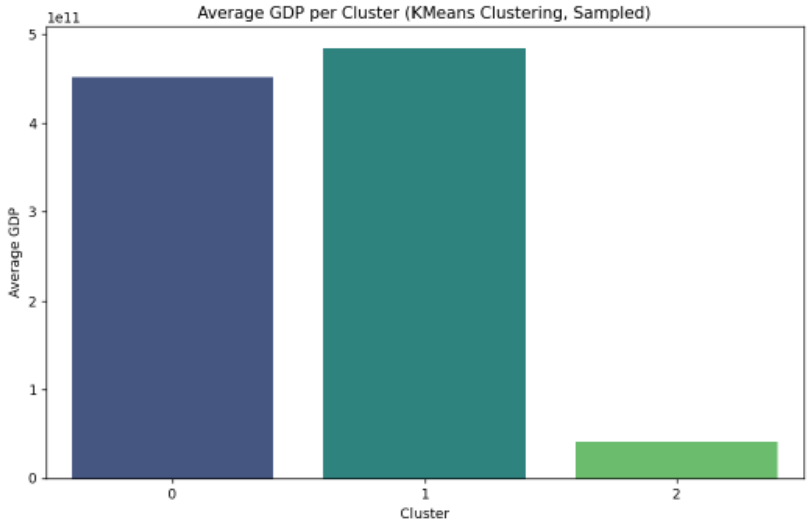






- **Cluster 0:** include paesi occidentali e mediorientali;
- **Cluster 1:** include paesi in via di sviluppo e paesi sviluppati come Asia e Europa orientale;
- **Cluster 2:** include paesi prevalentemente africani evidenziando un minore sviluppo economico.

Anche in questo caso sono stati introdotti nuovi dati e fattori nell'analisi, in particolari il GDP.



CONCLUSIONI

L'analisi ha evidenziato disparità economiche tra i cluster, evidenziando che il GDP può influenzare i risultati dei disturbi mentali, infatti, le nazioni con GDP più alto tendono a essere raggruppate insieme e questo è importante perché serve a comprendere come il fattore economico è molto importanti nella prevalenza dei disturbi mentali.
Sia nei cluster trovati dal KMeans e quelli dall'agglomerative clustering:

- Le nazioni con un GDP alto mostrano maggiore prevalenza nei disturbi mentali;
- Le nazioni con un GDP basso potrebbero avere una prevalenza sottostimata dei disturbi mentali a causa di limitazioni nelle risorse diagnostiche.

Metodologia	**Numero di cluster**	**Coefficiente di Silhouette**	**Descrizione dei Cluster**
Clustering Agglomerative	3	0.388	Cluster ragionevolmente separati ma potrebbe esserci qualche sovrapposizione
Kmeans Clustering	3	0.400	Cluster ragionevolmente separati ma potrebbe esserci qualche sovrapposizione

PROFILI INDIVIDUATI:

	Descrizione	PIL (medio)	Prevalenza di Disturbi	Proposte di interventi
Cluster 0	Include nazioni economicamente diverse con uno sviluppo da moderato a elevato, unendo paesi occidentali e mediorientali. Tra cui: Germania, Regno Unito,Francia,Italia, Canada,Paesi Bassi,Arabia Saudita, Emirati Arabi Uniti	$42.21 miliardi	Alta prevalenza di disturbi depressivi e d'ansia. Bassa prevalenza di disturbi alimentari	Potenziamento delle infrastrutture sanitarie. Campagne di sensibilizzazione per ridurre lo stigma. Supporto ai familiari dei pazienti con disturbi mentali.
Cluster 1	Include un ampio mix di paesi in via di sviluppo e sviluppati, prevalentemente dell'Asia e dell'Europa orientale. Tra cui:Cina India, Russia, Brasile, Sud Africa, Turchia, Polonia, Indonesia, Malesia,	$48.25 miliardi	Alta prevalenza di disturbi depressivi.	Accesso equo ai servizi di salute mentale e sensibilizzazione

	Ucraina			
Cluster 2	Prevalentemente composto da paesi africani.	$4.84 miliardi	Alta prevalenza di disturbi d'ansia e bipolari. Prevalenza di schizofrenia e disturbi alimentari relativamente alti	Sviluppo di politiche di salute mentale nazionali. Rafforzamento delle capacità di diagnosi e trattamento. Programmi di prevenzione e promozione della salute mentale.

PROPOSTE DI PIANI DI INTERVENTO:

Sulla base delle raccomandazioni dell'Organizzazione Mondiale della Sanità (OMS), è possibile individuare alcune strategie di intervento mirate per i principali disturbi mentali analizzati.

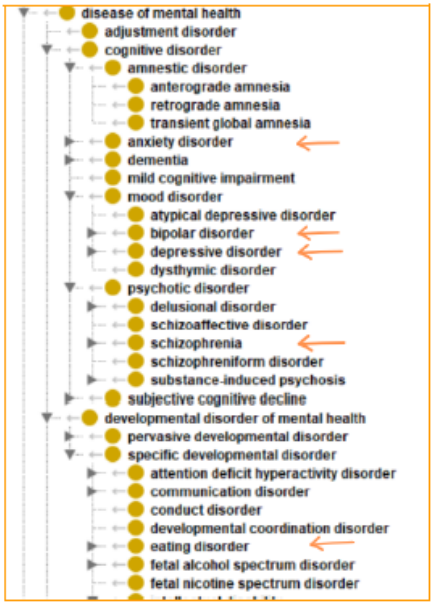
- **Schizofrenia**
  Per la schizofrenia risulta fondamentale promuovere la diagnosi precoce e l'avvio tempestivo del trattamento, al fine di ridurre la gravità dei sintomi e migliorare la qualità della vita del paziente nel lungo periodo.
- **Depressione e disturbi d'ansia**
  Per la depressione e i disturbi d'ansia è particolarmente importante implementare programmi di sensibilizzazione rivolti alla popolazione, finalizzati al riconoscimento precoce dei sintomi e alla diffusione di strategie di gestione dello stress.
  Inoltre, è essenziale sviluppare programmi di riabilitazione e supporto sociale che favoriscano il reinserimento del paziente nella società e riducano il rischio di isolamento e stigmatizzazione.

ONTOLOGIA

Per la gestione e l'integrazione dei dati è stato adottato un approccio ontologico, supportato da diversi strumenti software.
In particolare, **pandas** è stato utilizzato per il caricamento e la manipolazione dei dataset, **rdflib** per la creazione e l'interrogazione dei grafi RDF, mentre **Protégé** è stato impiegato per l'esplorazione e l'analisi della struttura dell'ontologia.
È stata progettata un'ontologia dedicata con lo scopo di organizzare in modo strutturato i dati e i risultati ottenuti, facilitandone la condivisione, il riutilizzo e l'integrazione con altre fonti informative.
L'ontologia sviluppata è stata successivamente integrata con la **Human Disease Ontology**, permettendo l'utilizzo di definizioni e relazioni standardizzate già esistenti e garantendo così una maggiore interoperabilità dei dati.

HUMAN DISEASE ONTOLOGY

*La **Human Disease Ontology (HDO)** rappresenta un'ontologia standardizzata per la descrizione delle malattie umane, progettata per fornire un vocabolario medico coerente, riutilizzabile e condiviso all'interno della comunità biomedica.*
Essa descrive le patologie umane, le relative caratteristiche fenotipiche e i concetti associati, ed è sviluppata attraverso un processo collaborativo che coinvolge ricercatori e istituzioni accademiche, tra cui la Scuola di Medicina dell'Università del Maryland e l'Istituto per le Scienze del Genoma.
*In una fase preliminare è stata analizzata la struttura dell'ontologia tramite **Protégé**, al fine di comprenderne l'organizzazione gerarchica.*
*La classe **Disease of mental health** rappresenta il punto di riferimento per le patologie considerate nello studio ed è definita come sottoclasse della classe **Disease**.*
All'interno di essa, ciascuna patologia principale è ulteriormente articolata in sottoclassi, consentendo una rappresentazione dettagliata e coerente dei disturbi mentali analizzati.



Schizophrenia è suddivisa in diverse forme, ciascuna con caratteristiche distintive.

Depression include vari disturbi depressivi con diverse intensità e durata.

Bipolar Disorders comprende vari tipi di disturbo bipolare, con differenze significative negli episodi di mania e depressione.

Eating Disorders copre vari disturbi alimentari, con differenti manifestazioni comportamentali e psicologiche.

Anxiety Disorders include diversi tipi di disturbi d'ansia, con una gamma di sintomi che spaziano dall'ansia generalizzata alle fobie specifiche.

METODOLOGIA:
Per prevenire conflitti semantici e favorire l'integrazione con altre ontologie o dataset esterni, sono stati definiti due namespace distinti.
Il namespace **PREDICT** è stato utilizzato per l'ontologia personalizzata sviluppata all'interno del progetto, mentre il namespace **OBO** è stato adottato per il collegamento con l'Ontologia delle Malattie Umane.
I disturbi mentali presenti nel dataset sono stati associati agli URI (Uniform Resource Identifier) corrispondenti definiti nella Human Disease Ontology.
Per esempio, la schizofrenia è stata mappata all'URI: https://disease-ontology.org/?id=DOID:5419, i disturbi depressivi all'URI: http://purl.obolibrary.org/obo/DOID_1596, e così via.
È stato quindi costruito un grafo RDF per rappresentare le informazioni, composto da nodi (individui) e archi (relazioni) che descrivono i collegamenti tra gli elementi del dominio.
Sono state definite apposite proprietà per mettere in relazione i disturbi mentali con le nazioni presenti nel dataset.
Per ogni record del dataset sono state generate triple RDF che collegano i paesi (soggetti) agli attributi associati (predicati e oggetti), quali l'anno di riferimento e le tipologie di disturbi mentali.
Infine, il grafo è stato serializzato in un file OWL (*IntegratedOntology.owl*). La serializzazione consente di trasformare una struttura dati in un formato facilmente memorizzabile; in particolare, il formato OWL permette la rappresentazione di ontologie complesse.

CONCLUSIONI FINALI

Le analisi condotte a livello globale mostrano che la diffusione dei disturbi mentali varia significativamente tra le diverse regioni del mondo ed è fortemente influenzata da fattori culturali, economici e, soprattutto, dall'accesso ai servizi sanitari.

Inoltre, attraverso l'applicazione di tecniche di apprendimento non supervisionato integrate con il dato sul GDP, è stato possibile osservare come le disuguaglianze economiche tra i paesi incidano sulla prevalenza e sulla gestione dei disturbi mentali.

Le previsioni future dei DALYs per i disturbi mentali in Italia, ottenute utilizzando il modello Random Forest, indicano una tendenza complessiva all'aumento, in particolare per schizofrenia, depressione e disturbi d'ansia. Questo risultato evidenzia l'importanza di strategie di prevenzione e intervento all'interno della società.