Università degli Studi di Bari "Aldo Moro"
Laurea Magistrale in Data Science

Progetto Data Mining

Emergency Events Database

Anisa Bakiu
Matricola: 837632

EM-DATA (Emergency Events Database), una risorsa di riferimento per lo studio dei disastri naturali e tecnologici e dei loro impatti a livello globale.

Anno Accademico 2025/2026

Indice

1 Data Understanding

1.1 Descrizione del dataset

Nel 1988, il Centro di Ricerca sull'Epidemiologia dei Disastri (CRED), presso l'Università di Lovanio (UCLouvain), ha lanciato l'Emergency Events Database (EM-DAT). Il database EM-DAT registra le catastrofi di massa e il loro impatto sanitario ed economico a livello nazionale.
L'obiettivo iniziale del database è quello di supportare l'azione umanitaria a livello nazionale e internazionale.
Oggi, l'EM-DAT viene utilizzato anche per razionalizzare la preparazione alle catastrofi e il processo decisionale, fornendo al contempo una base oggettiva per la valutazione della vulnerabilità e del rischio.
Il database contiene dati essenziali sull'occorrenza e gli effetti di 27.000 disastri in tutto il mondo dal 1900 a oggi, ed è compilato da diverse fonti di informazione, tra cui agenzie delle Nazioni Unite, organizzazioni non governative, compagnie assicurative, istituti di ricerca e agenzie di stampa [1].
Formalmente, la definizione di disastro dell'EM-DAT è:

> **Disastro**: Una situazione o un evento che supera le capacità locali, rendendo necessaria una richiesta di assistenza esterna a livello nazionale o internazionale; un evento imprevisto e spesso improvviso che causa gravi danni, distruzione e sofferenza umana.

L'EM-DAT cataloga solo i disastri che soddisfano i suoi criteri di inclusione, e registra sia i disastri causati da calamità naturali che quelli tecnologici. Questi ultimi sono incidenti non intenzionali e non situazioni di conflitto, violenza o terrorismo [2].
L'albero di classificazione principale presenta 4 livelli di profondità, pertanto i disastri sono suddivisi in gruppi, sottogruppi, tipi e sottotipi, come presentato nella tabella successiva. I due gruppi di disastri EM-DAT sono "Naturali" e "Tecnologici".

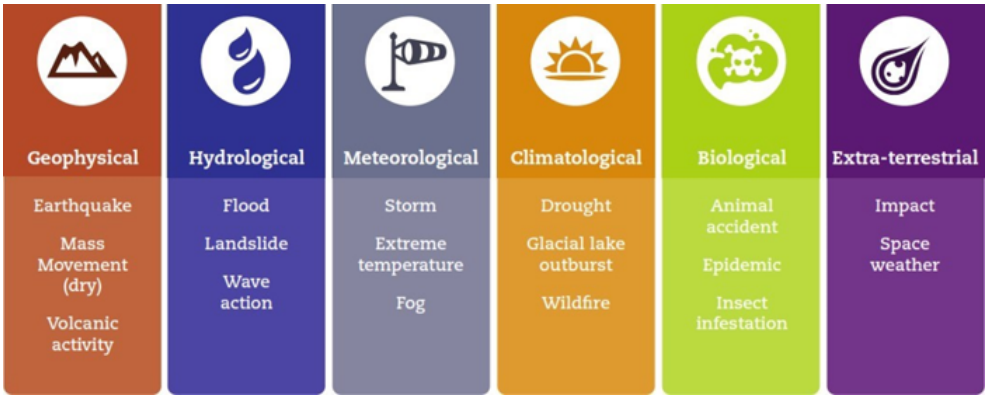

Figure 1: Classificazione dei disastri

1.2 Descrizione dei dati

Dal seguente output in Python, è possibile capire le variabili presenti nel database e comprendere la struttura delle informazioni disponibili per le analisi.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27262 entries, 0 to 27261
Data columns (total 46 columns):
#   Column                                    Non-Null Count  Dtype
---  ------                                    --------------  -----
0   DisNo.                                    27262 non-null  object
1   Historic                                  27262 non-null  object
2   Classification Key                        27262 non-null  object
3   Disaster Group                            27262 non-null  object
4   Disaster Subgroup                         27262 non-null  object
5   Disaster Type                             27262 non-null  object
6   Disaster Subtype                          27262 non-null  object
7   External IDs                              4338 non-null   object
8   Event Name                                8579 non-null   object
9   ISO                                       27262 non-null  object
10  Country                                   27262 non-null  object
11  Subregion                                 27262 non-null  object
12  Region                                    27262 non-null  object
13  Location                                  24758 non-null  object
14  Origin                                    4826 non-null   object
15  Associated Types                          4220 non-null   object
16  OFDA/BHA Response                         27262 non-null  object
17  Appeal                                    27262 non-null  object
18  Declaration                               27262 non-null  object
19  AID Contribution ('000 US$)               787 non-null    float64
20  Magnitude                                 5226 non-null   float64
21  Magnitude Scale                           17085 non-null  object
22  Latitude                                  2817 non-null   float64
23  Longitude                                 2817 non-null   float64
24  River Basin                               1615 non-null   object
25  Start Year                                27262 non-null  int64
26  Start Month                               26774 non-null  float64
27  Start Day                                 23291 non-null  float64
28  End Year                                  27262 non-null  int64
29  End Month                                 26484 non-null  float64
30  End Day                                   23397 non-null  float64
31  Total Deaths                              21740 non-null  float64
32  No. Injured                               8999 non-null   float64
33  No. Affected                              11470 non-null  float64
34  No. Homeless                              2701 non-null   float64
35  Total Affected                            18387 non-null  float64
36  Reconstruction Costs ('000 US$)           42 non-null     float64
37  Reconstruction Costs, Adjusted ('000 US$) 42 non-null     float64
38  Insured Damage ('000 US$)                 1151 non-null   float64
39  Insured Damage, Adjusted ('000 US$)       1144 non-null   float64
```

```
40  Total Damage ('000 US$)                  5898 non-null   float64
41  Total Damage, Adjusted ('000 US$)        5875 non-null   float64
42  CPI                                      26912 non-null  float64
43  Admin Units                              8415 non-null   object
44  Entry Date                               27262 non-null  object
45  Last Update                              27262 non-null  object
dtypes: float64(20), int64(2), object(24)
memory usage: 9.6+ MB
```

1.3 Flusso del lavoro

Nella figura 2 presento il flusso del lavoro di questo progetto.
Il database EM-DATA è composto da 27.262 righe e 46 colonne, pertanto iniziamo con un'analisi esplorativa per comprenderne meglio il contenuto.
Dall'analisi emerge una notevole quantità di dati mancanti. Per questo motivo, ho cercato di scoprire quale ipotesi di dati mancanti si verifica in questo database (MCAR, MAR o MNAR), in modo da applicare la tecnica di imputazione più appropriata all'ipotesi confermata dai test statistici.
Dopo aver imputato i dati, procedo quindi all'analisi di clustering e all'estrazione delle regole di associazione. I parametri delle diverse tecniche sono scelti in base alle prestazioni ottenute, non in modo arbitrario.

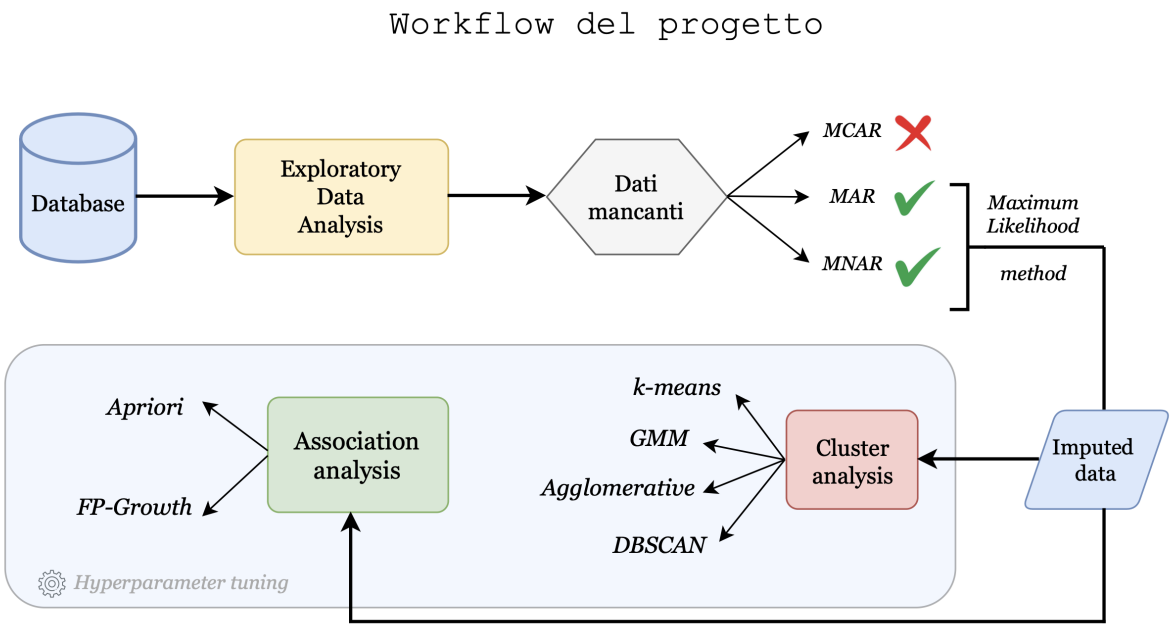

Figure 2: Flusso del lavoro

1.4 Statistiche descrittive

Le analisi e le varie visualizzazioni che seguono, aggregano i dati e mettono in evidenza i pattern che emergono, senza ricorrere ancora a elaborazioni complesse.

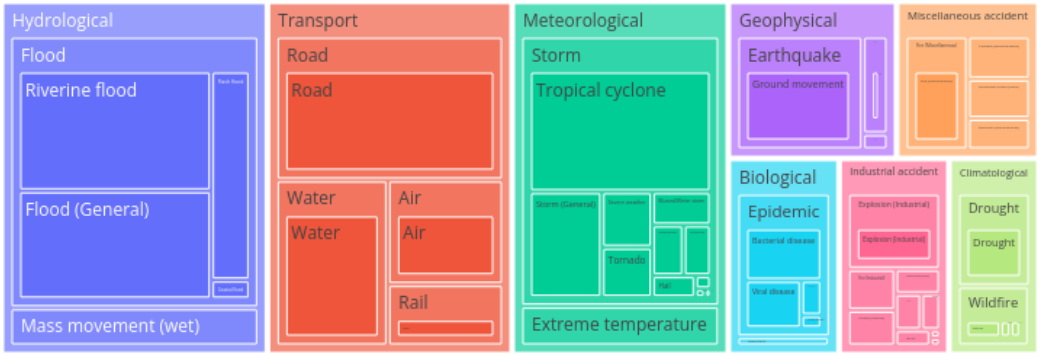

Figure 3: Gerarchia dei disastri nel database

L'obiettivo di questo grafico è comprendere come i diversi disastri siano organizzati gerarchicamente all'interno del database. Utilizzando una `treemap`, possiamo visualizzare in modo chiaro le relazioni tra sottogruppi, tipi e sottotipi di disastri, evidenziando eventuali concentrazioni o pattern interessanti.
Dal grafico emerge che le tre categorie con il maggior numero di disastri sono: *Hydrological* con 6991 disastri, *Transport* con 6431 e *Meteorological* 5681 disastri.

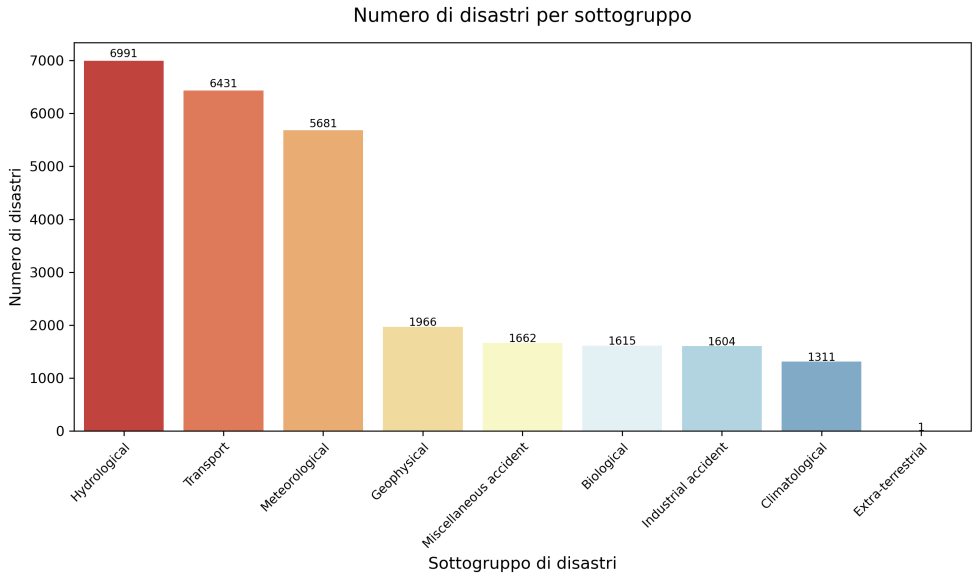

Figure 4: Numero di disastri per sottogruppo

L'obiettivo di questa rappresentazione è analizzare la distribuzione dei disastri per sottogruppo. Si osserva chiaramente che i sottogruppi Hydrological, Transport e Meteorological sono quelli più rappresentati nel database, mentre un numero significativamente minore di eventi è associato ai disastri di tipo Geophysical.

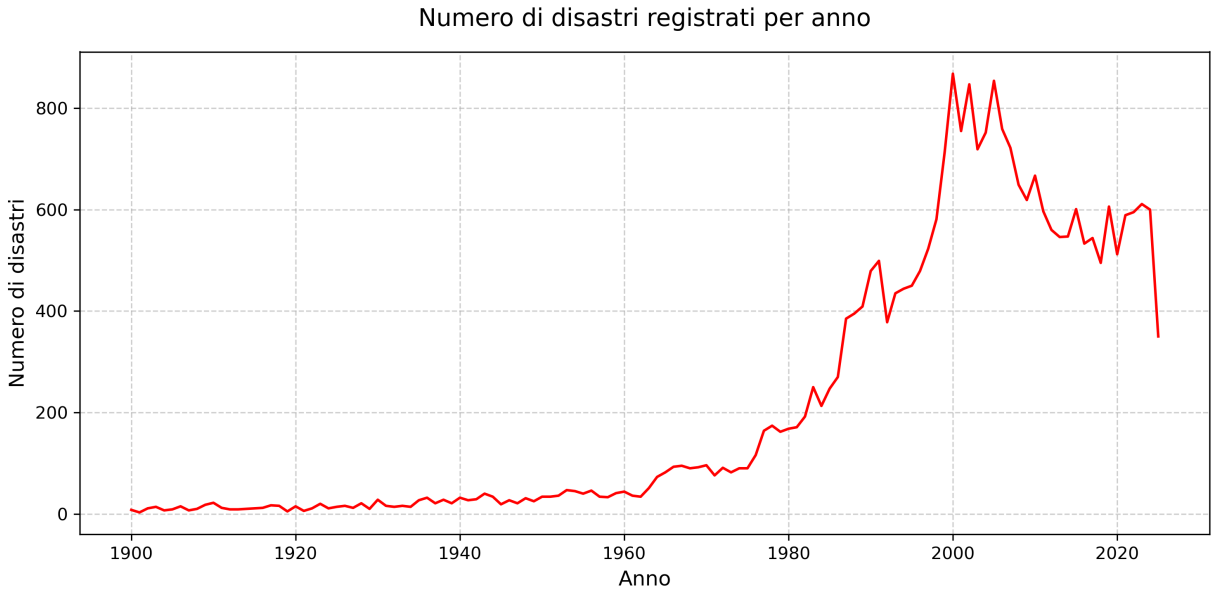

Figure 5: Numero di disastri registrati per anno

Da questo grafico emerge un picco nel numero di eventi registrati intorno agli anni 2000, seguito da una lieve diminuzione agli inizi degli anni 2020. Si osserva inoltre che, fino al 1980, il numero di disastri segnalati risulta molto limitato, ciò potrebbe essere dovuto alla mancanza di strumenti per la raccolta dei dati. Questa ipotesi sarà approfondita nella sezione dedicata all'analisi dei dati mancanti.

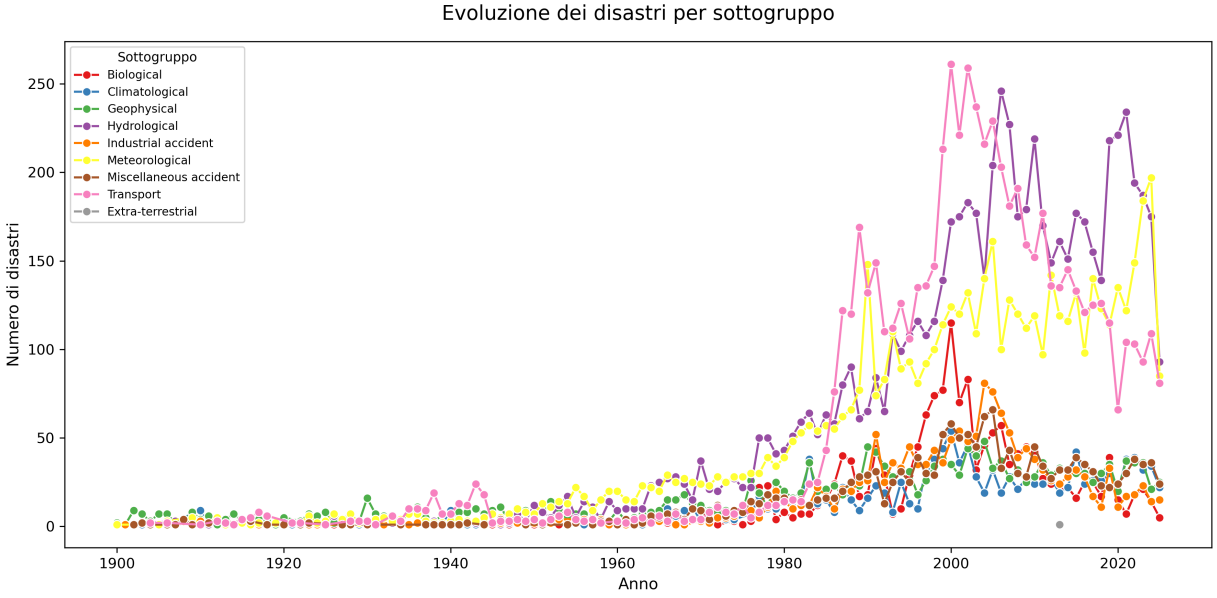

Figure 6: Evoluzione dei disastri per sottogruppo

Dal grafico si osserva che il picco di disastri registrato negli anni 2000 è principalmente dovuto alle sottocategorie di *Transport*, *Hydrological* e *Biological*. Si nota inoltre un incremento significativo dei disastri di tipo *Meteorological*, che contribuiscono in misura rilevante alla variazione complessiva di questo periodo.

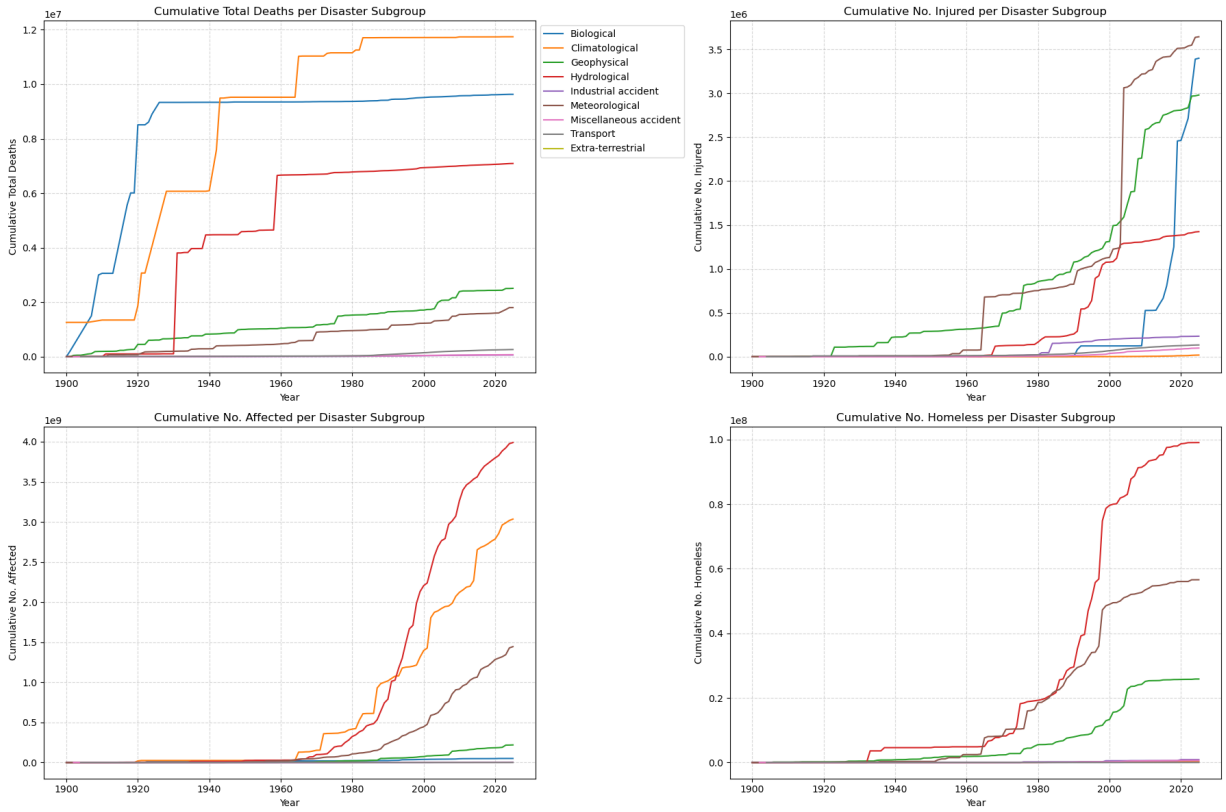

Figure 7: Andamento cumulativo degli impatti dei disastri nel tempo, suddivisi per sottogruppo

L'analisi esplorativa cumulativa fornisce una visione più trasparente dei pattern storici e può supportare ulteriori analisi.
Dal database emerge che il numero di decessi dovuti a cause climatologiche è significativamente più alto rispetto agli altri sottogruppi e mostra un incremento ciclico ogni venti anni, con l'ultimo aumento registrato intorno al 2000. Seguono i decessi per cause biologiche, che hanno registrato l'ultima crescita significativa negli anni '20 del secolo scorso.
Per quanto riguarda il numero di persone ferite a seguito di disastri, i maggiori incrementi si osservano nei sottogruppi meteorologici e geofisici, mentre le cause biologiche mostrano un aumento meno pronunciato dopo il 2000.
I grafici relativi al numero di colpiti e di sfollati, presentano trend simili. Nel primo, si nota un incremento marcato degli individui colpiti da disastri idrologici e climatologici, mentre nel secondo grafico, il secondo sottogruppo è quello meteorologico.
Analizzando l'andamento dei grafici relativi ai colpiti e ai sfollati, si evidenzia che le cause meteorologiche sono responsabili della maggior parte degli individui rimasti feriti o sfollati.

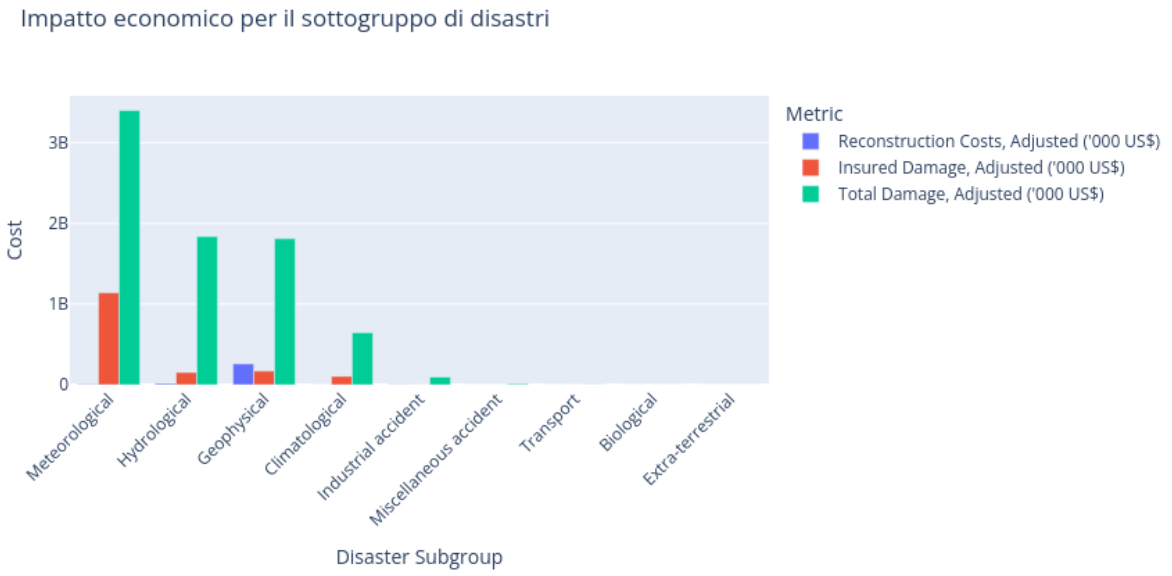

Figure 8: Impatto economico per sottogruppo

Dall'impatto economico per sottogruppo, emerge che i disastri meteorologici sono quelli che causano i danni maggiori, seguiti dai disastri idrologici e geofisici, considerando il totale dei danni normalizzati tramite l'indice dell'inflazione CPI (Consumer Price Index).
Si osserva inoltre che, tra tutti i sottogruppi, solo per i disastri geofisici è presente una somma rilevante destinata alla ricostruzione, evidenziando un comportamento differente nella gestione dei fondi.

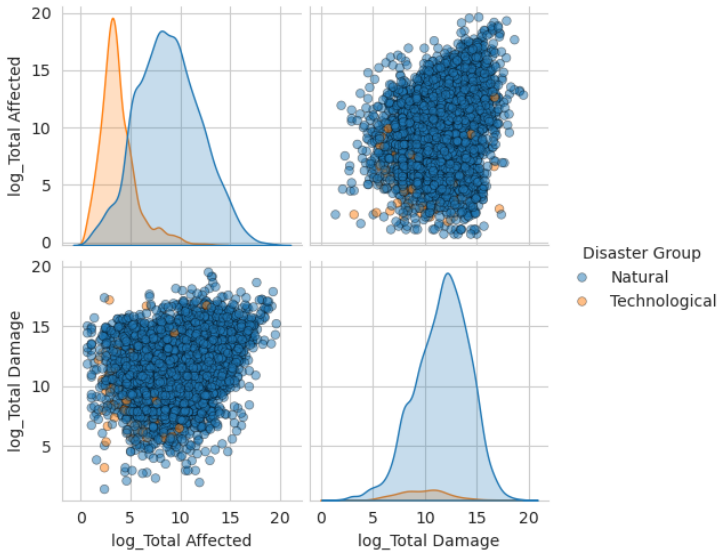

Figure 9: Distribuzione di danni economici per gruppo di disastro

La distribuzione del numero di persone colpite dai disastri tecnologici (in arancione) è unimodale e leggermente asimmetrica a sinistra, mentre la distribuzione delle persone colpite dai disastri naturali (in blu) mostra una variabilità più elevata.
Si osserva inoltre una differenza significativa nella distribuzione della variabile relativa ai

danni economici totali. I danni causati dai disastri naturali risultano generalmente molto più elevati rispetto a quelli generati dai disastri tecnologici.

Infine, possiamo porci la seguente domanda: "I disastri più costosi ricevono più facilmente una dichiarazione di emergenza?"
Per rispondere, analizziamo la tabella (c.d. *"Heatmap"*) sottostante.

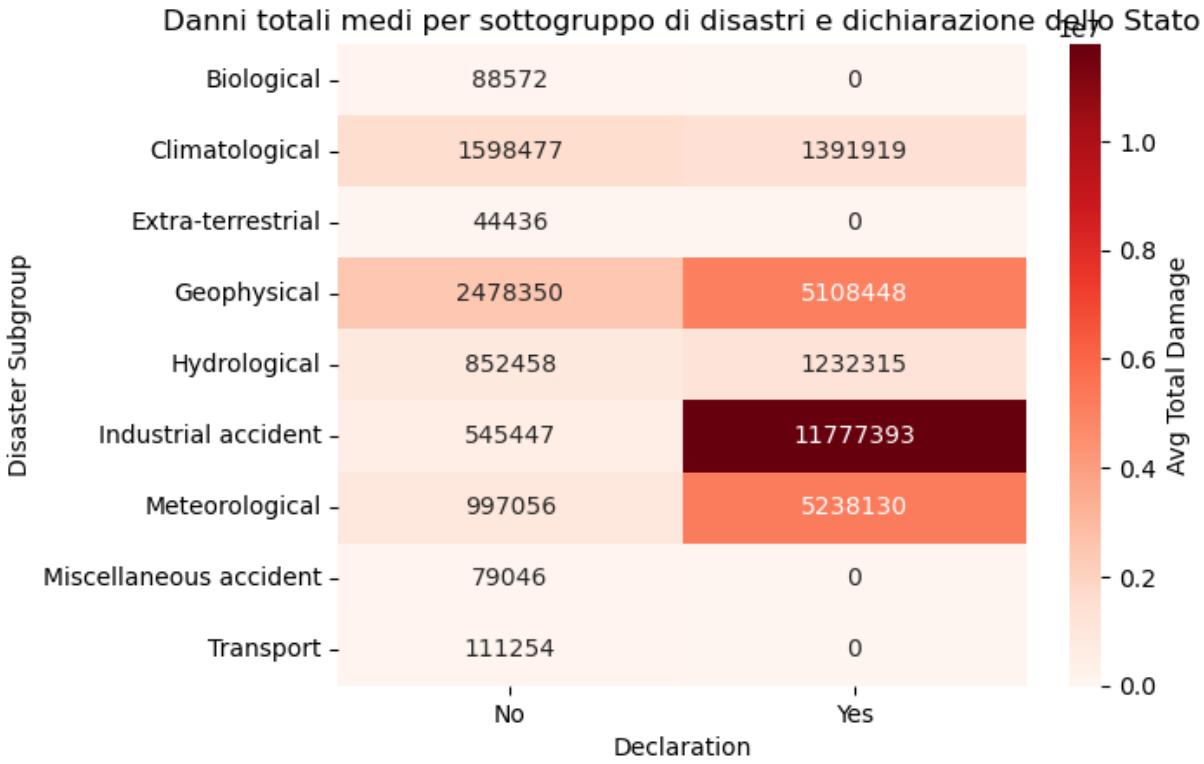

Figure 10: Danni totali medi per sottogruppo di disastri e dichiarazione dello Stato

Qui ho messo in relazione i sottogruppi di disastri con la risposta dello Stato e il danno economico medio. Ho usato come righe i diversi sottogruppi di disastri, come colonne il valore della variabile Declaration (Yes/No) e come valori numerici il danno totale medio per ciascun sottogruppo.
Un valore alto nella colonna "Yes", significa che quando lo Stato ha dichiarato emergenza per quel tipo di disastro, il danno medio registrato è elevato.
Dal heatmap emerge l'opportunità di individuare pattern molto interessanti. In particolare, i disastri che provocano i danni economici maggiori sono quelli in grado di generare situazioni di emergenza a livello statale. Tuttavia, si osserva che altri eventi, appartenenti alle categorie Geophysical, Climatological e Meteorological, pur causando danni significativi, non determinano necessariamente situazioni di emergenza.
Questa analisi evidenzia come *la semplice entità del danno economico non sia sufficiente per spiegare l'impatto complessivo dei disastri*, e sottolinea l'importanza di individuare pattern e correlazioni attraverso tecniche di Data Mining.

1.5 Rappresentazioni spaziali dei dati

In questa sezione, rappresento la distribuzione dei disastri nel mondo tramite mappe tematiche.

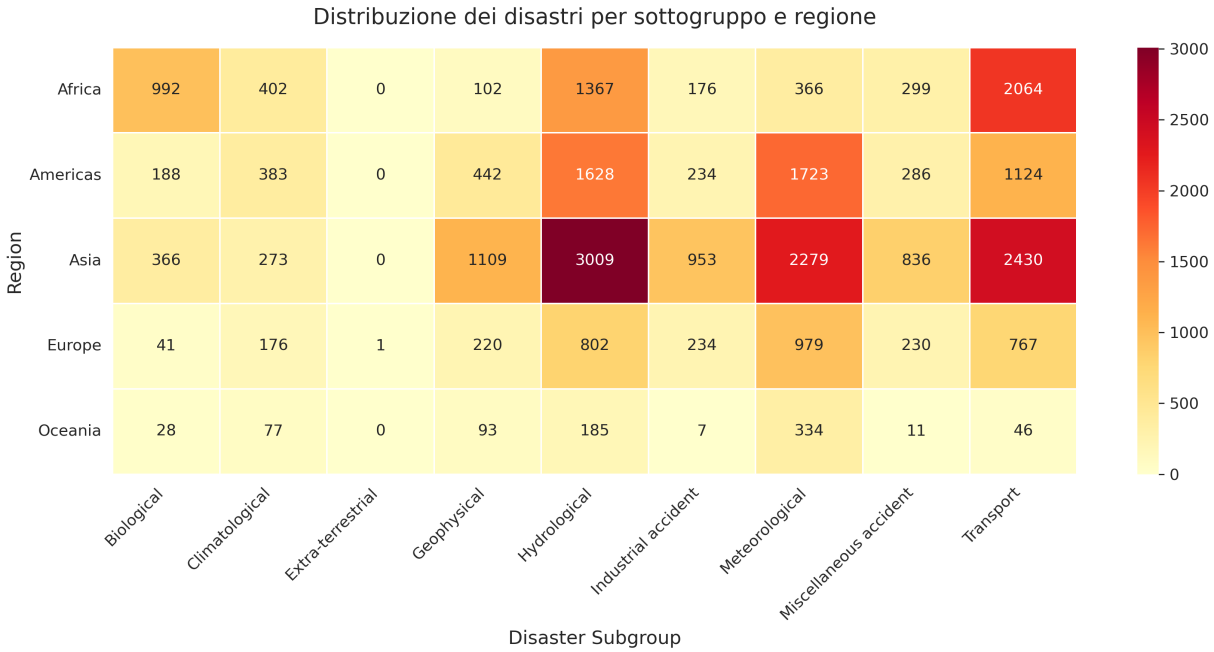

Figure 11: Distribuzione dei disastri per sottogruppo e regione

Dall'heatmap risultante, è possibile osservare quali regioni sono maggiormente colpite dai disastri e a quale sottogruppo appartengono.
Ad esempio, emerge chiaramente che l'Asia è il continente con il maggior numero di disastri, seguita da America e Africa.
In Asia, i disastri più frequenti sono di tipo idrologico, meteorologico e di trasporto, una tendenza che si riscontra anche in Africa e America, sebbene con valori inferiori.

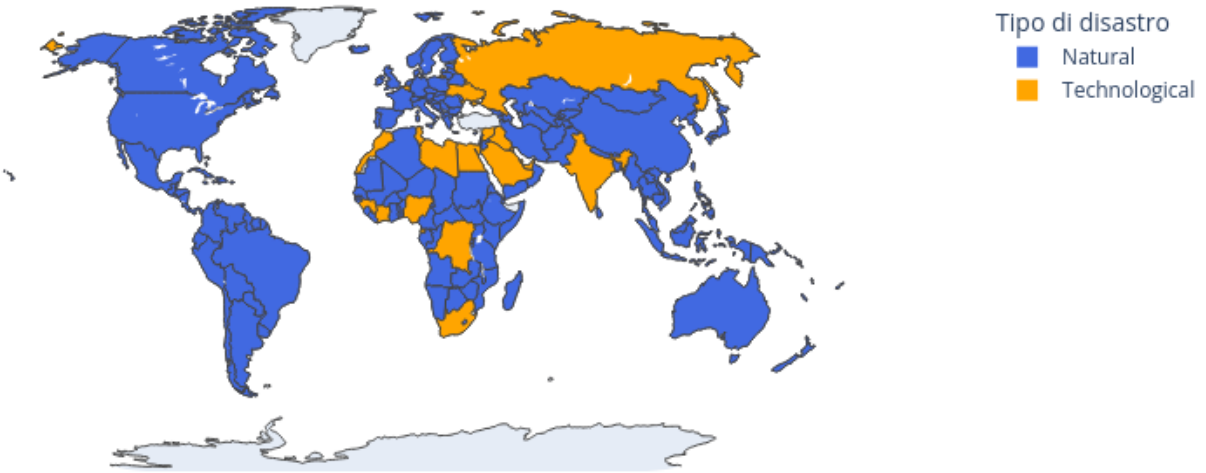

Figure 12: Gruppo di disastro predominante per paese.

I disastri di tipo naturale predominano a livello globale, con alcune eccezioni nei paesi asiatici, come Russia e India, e in alcune aree dell'Africa.

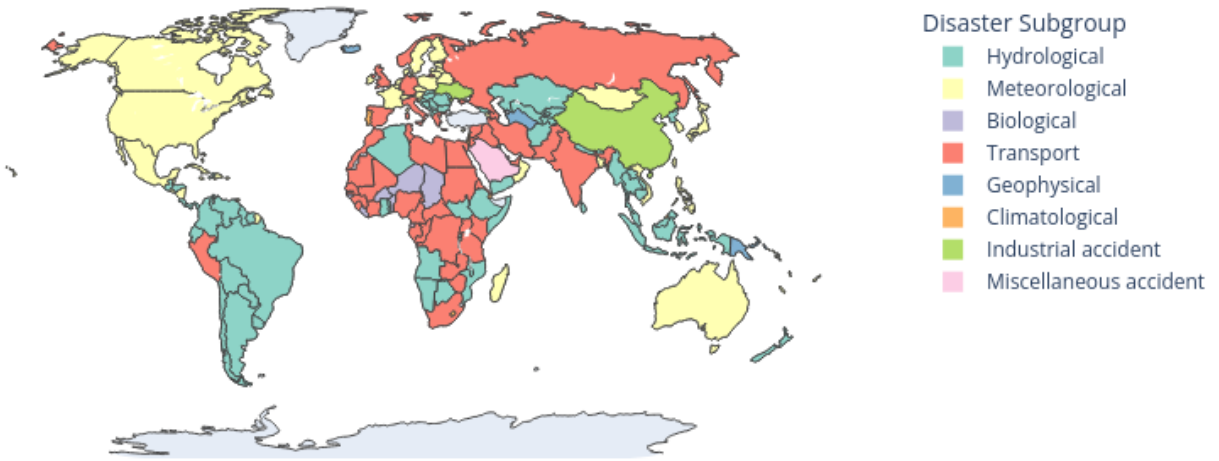

Figure 13: Sottogruppo di disastro per paese

Da questa analisi è possibile osservare in modo più dettagliato quale sottogruppo di disastri risulta più frequente in ciascun paese.

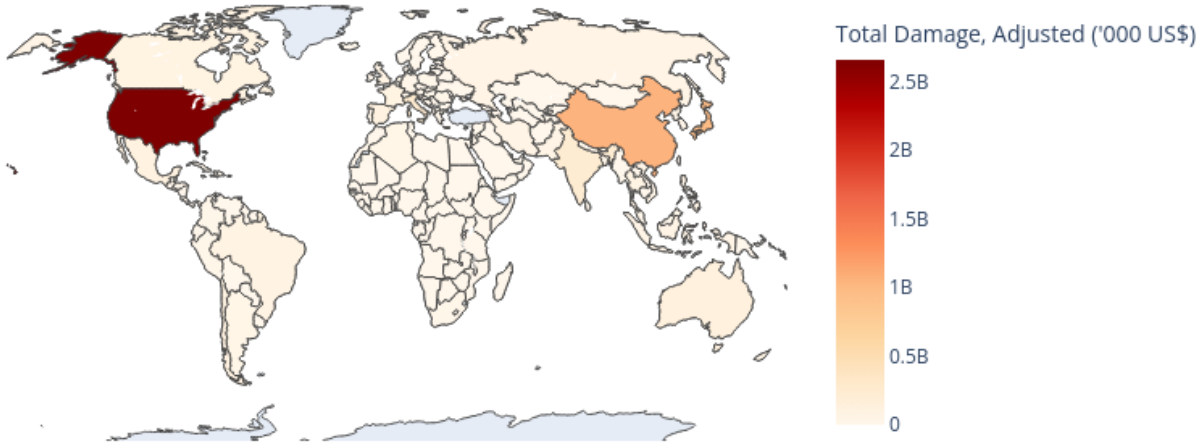

Figure 14: Danni totali per paese

Dalla mappa emerge chiaramente che i disastri negli Stati Uniti e in Cina sono associati ai danni economici più elevati.

2 Data Preparation

2.1 Selezione delle variabili

Considerato l'elevato numero di variabili presenti nel database, è fondamentale valutarne l'importanza mediante le tecniche di Data Mining studiate, al fine di includere nelle analisi successive solo quelle che risultano significativamente rilevanti.

2.1.1 Costruzione del dato

Prima di procedere, è necessario creare mediante una trasformazione, una nuova variabile non presente originariamente nel database.
Sul sito ufficiale si segnala che, per i disastri che si sviluppano gradualmente nel tempo, il giorno e il mese esatti potrebbero non essere sempre riportati [3].
Per questo motivo, le colonne contenenti queste informazioni sono state convertite in formato `datetime`, consentendo il calcolo della durata degli eventi.
Per valutare l'importanza delle variabili, si utilizza il calcolo del coefficiente di correlazione di Spearman.

2.1.2 Correlazione di Spearman

Table 1: Coefficienti di correlazione di Spearman tra le variabili selezionate e la variabile di interesse

Feature	Spearman ρ	p-value
Total Affected	-0.659539	0.000000e+00
duration_days	-0.426700	0.000000e+00
No. Affected	-0.355168	0.000000e+00
No. Injured	-0.188847	5.016955e-73
Total Damage, Adjusted ('000 US$)	-0.143467	2.158742e-28
Total Deaths	0.097825	2.242214e-47
No. Homeless	-0.081987	1.988638e-05
Magnitude	0.029926	3.051410e-02
CPI	-0.010234	9.319865e-02
Insured Damage, Adjusted ('000 US$)	-0.017455	5.553479e-01
Reconstruction Costs, Adjusted ('000 US$)	-0.036898	8.165491e-01

Il risultato mostra la correlazione di Spearman tra le variabili numeriche del database e la variabile `Disaster Group`. Il coefficiente di Spearman, indicato come `Spearman` ρ, misura quanto due variabili variano insieme in modo monotono.
Se il valore è vicino a `+1` la relazione è fortemente positiva, se è vicino a `-1` è fortemente negativa, mentre valori prossimi a 0 indicano che non c'è una relazione monotona rilevante.
In questo caso, i valori negativi indicano che passando da un gruppo di disastri a un altro, la variabile tende a diminuire.
Il valore di `p-value` serve a capire se la correlazione osservata è statisticamente significativa. Quando è minore di `0.05` possiamo considerare che la relazione non è casuale. Se invece è più alto, la correlazione non è affidabile.
Nel output, le variabili che mostrano correlazioni significative con il tipo di disastro sono:

1. Total Affected

2. duration days
3. No Affected
4. No. Injured
5. Total Damage, Adjusted ('000 US $)
6. Total Death
7. No. Homeless
8. Magnitude

Chiaramente, il numero totale di persone colpite e la durata dell'evento variano in modo consistente tra i diversi gruppi di disastri, fornendo quindi informazioni utili.
Al contrario, variabili come "CPI", "Insured Damage" o "Reconstruction Costs" presentano correlazioni molto deboli e p-value alti, quindi non sembrano fornire informazioni rilevanti per discriminare i gruppi di disastro.
Ho utilizzato la correlazione di Spearman perché è una misura non parametrica, cioè non richiede che i dati seguano una distribuzione normale. Nel database EM-DAT molte variabili numeriche, come "Total Deaths" o "Total Affected", sono fortemente asimmetriche, con la maggior parte dei valori molto piccoli e pochi casi estremamente grandi.
In queste situazioni, la correlazione di Pearson non sarebbe adatta, perché è sensibile agli outlier e presume linearità.
Spearman invece, si basa sui ranghi dei valori, quindi riduce l'effetto dei valori estremi e misura quanto le variabili cambiano insieme in modo coerente (anche se non in modo lineare).

2.2 Gestione dei dati mancanti

Come riportato nella documentazione ufficiale, questo set di dati contiene molti dati mancanti [4].

> *For disaster events attributed to natural hazards occurring between 1990 and 2020, proportions of missing data on the human impacts of a disaster event were found to range from 1.3% - 22.3%. Proportions of missing data were much greater on the economic impacts, ranging from 41.5% - 96.2%.*

I dati non sono mancanti in modo casuale (Missing Completely At Random), ma nascondono schemi che potrebbero spiegare possibili relazioni, come indicato sotto:

> *The probability of missingness on the variables: number of people affected, number of deaths and total estimated damages (in US$) were partially explained by observed predictors of missingness: disaster type, income status of the country, disaster severity and the year the disaster occurred. For this reason, such missing data are unlikely to be missing completely at random (MCAR).*

Quindi, in questo caso, i metodi di gestione dei dati mancanti che si basano sul presupposto di MCAR sono inappropriati e potrebbero distorcere i risultati dello studio.

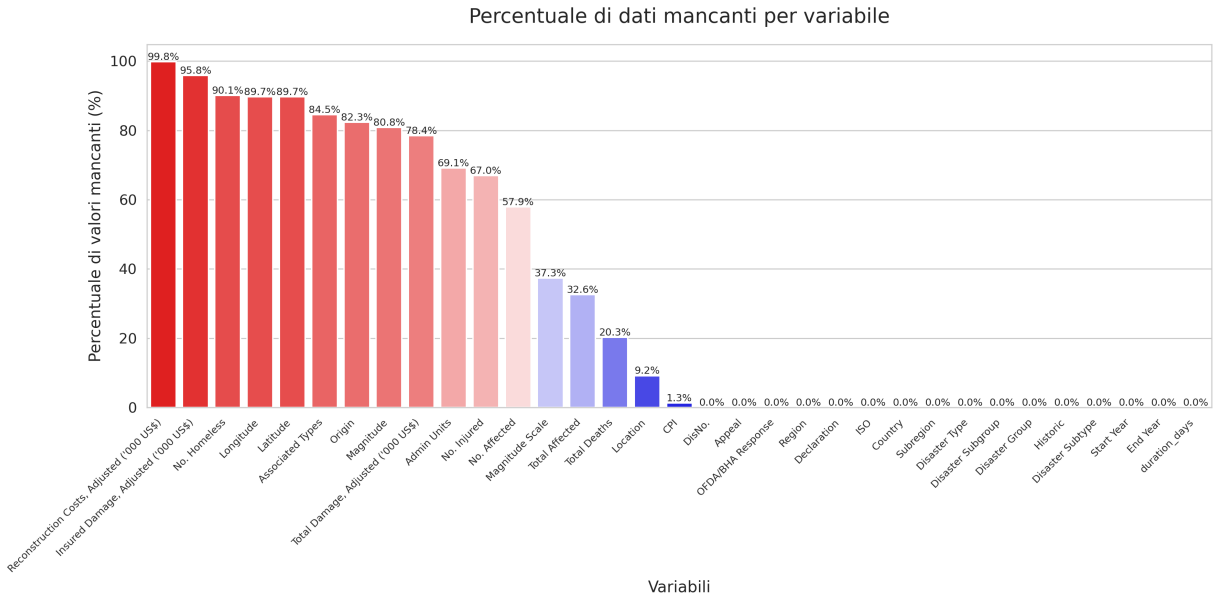

Figure 15: Percentuale di dati mancanti per variabile.

Dall'output sopra riportato possiamo notare che le prime 11 variabili hanno superato la soglia critica indicata nella documentazione del 60% di dati mancanti [4].
I metodi ampiamente utilizzati di imputazione dei dati, come k-NN o media/mediana, sono esplicitamente errati in quanto rappresentano una soluzione della categoria *MCAR*.
Inoltre, questi dati descrivono eventi critici che non possono essere dedotti da eventi simili in altri stati a causa della loro unicità.
In questa fase dell'analisi cercherò di considerare un modo più complesso per gestire questo problema. Indagherò il pattern sottostante e cercherò di comprendere la causa di tali dati mancanti.

2.2.1 Ipotesi MAR

La mia ipotesi è che la probabilità di mancata rilevazione delle variabili come numero di persone colpite, numero di morti e danni totali stimati sia spiegata da fattori osservati, quali il tipo di disastro, l'anno in cui si è verificato il disastro e da fattori non osservati, ad esempio lo stato di reddito del paese.
In questa sezione cercherò di trovare la correlazione di Pearson tra il numero di dati mancanti e le variabili indicate.

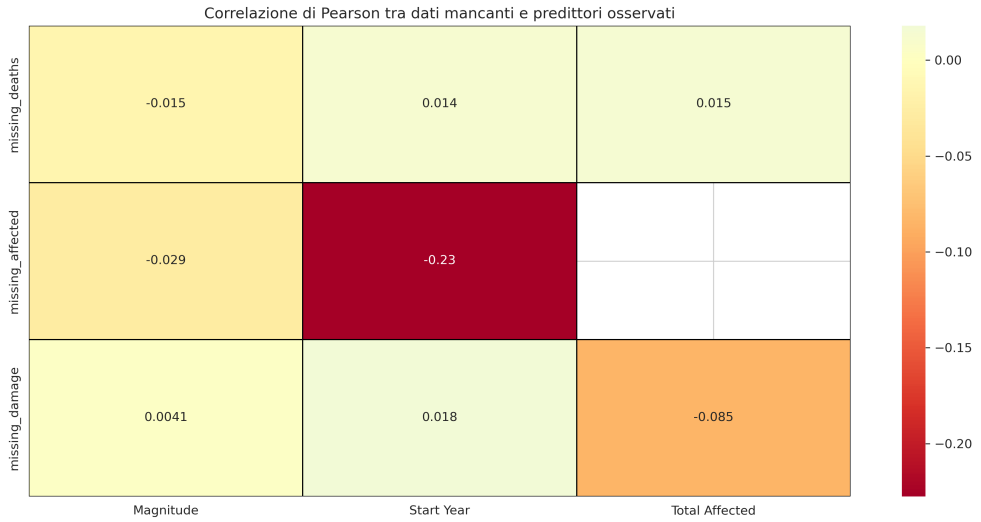

Figure 16: Correlazione di Pearson tra dati mancanti e predittori osservati

Notiamo che il valore di correlazione più basso disponibile è quello tra l'anno di inizio del disastro e le persone colpite.
La correlazione di -0.23 indica che i disastri più recenti hanno leggermente meno dati mancanti sul numero di persone colpite. È una correlazione debole ma coerente con l'idea che i dati storici sono meno completi.

2.2.2 Ipotesi MNAR

Se in MAR, la probabilità che un valore sia mancante dipende da altre variabili osservate nel set di dati, in MNAR, dipende da fattori non osservati, rendendolo il tipo più difficile da gestire.
Una variabile che non possiamo osservare in questo database è lo stato di reddito del paese/regione. Pertanto, la mia ipotesi è che i paesi più poveri sono meno in grado di segnalare perdite economiche. L'imputazione semplice tramite media e mediana, in tali condizioni produrrebbe stime altamente inaffidabili e distorte.

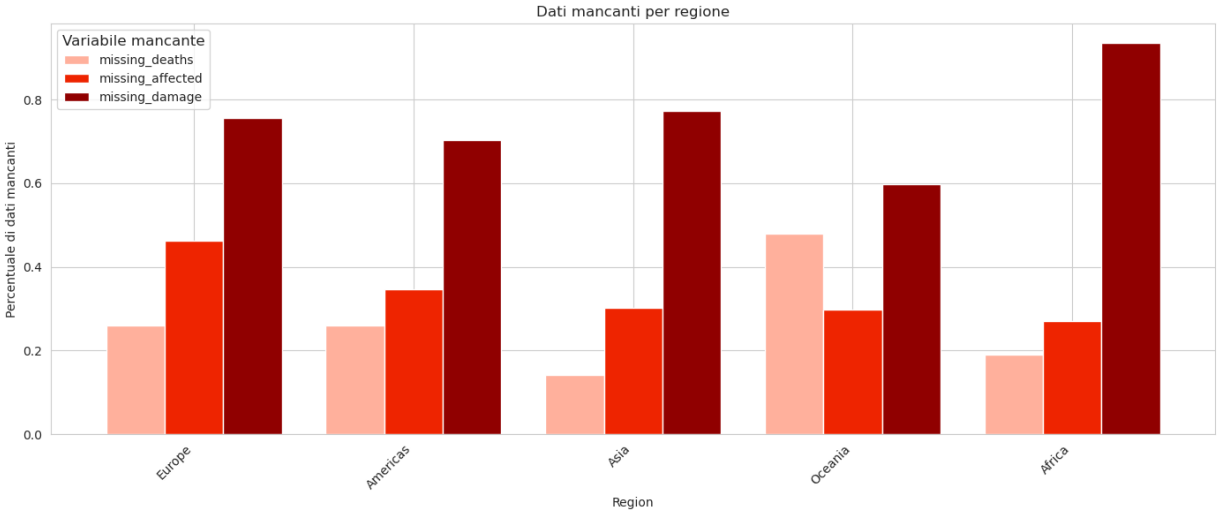

Figure 17: Dati mancanti per regione

Da questo grafico vediamo che l'Africa ha il numero più alto di dati mancanti nei danni economici. Ciò dimostra la teoria secondo cui l'Africa, con un PIL e risorse economiche relativamente bassi, non è stata in grado di segnalare (o valutare) i danni economici.
Per scoprire se la nostra ipotesi è anche statisticamente significativa, farò un test del "chi quadrato" e troverò il valore di "p".
L'idea è quella di verificare se la probabilità di avere dati mancanti dipende dalla regione. Il test confronta le frequenze osservate con quelle attese sotto l'ipotesi nulla di indipendenza.
`Chi2 = 1188.00, p-value = 0.00`

Il risultato dimostra che la probabilità della relazione osservata sia dovuta al caso è praticamente nulla. Questo conferma formalmente che i dati mancanti sono strutturati e supporta l'argomento teorico che la mancanza di informazioni economiche è legata alla capacità di reporting dei paesi. Quindi, la presenza di dati mancanti nei danni economici dipende fortemente dalla regione.

Pertanto, si procede su due fronti per la gestione dei dati mancanti:

1. **Eliminazione**: vengono rimosse le righe con più del 70% di valori numerici mancanti e tutte le colonne con circa il 80-90% di valori mancanti, poiché un numero così elevato di dati assenti non consentirebbe un'analisi affidabile.

2. **Imputazione**: sulle colonne rilevanti, che presentano una percentuale di dati mancanti inferiore all'80-90%, si applica un metodo di imputazione appropriato sotto l'ipotesi di tipo MNAR.

Il metodo di imputazione è stato scelto in conformità con la documentazione ufficiale di EM-DAT per la gestione dei dati mancanti, ovvero il metodo della "Massima Verosimiglianza (Maximum Likelihood)" [4].
Questo approccio utilizza i dati osservati per stimare i parametri più verosimili, ossia quelli che hanno la massima probabilità di generare i dati disponibili. Le verosimiglianze vengono calcolate separatamente per le osservazioni complete e incomplete rispetto alle variabili di interesse. Il prodotto di tali verosimiglianze viene poi massimizzato per ottenere la stima dei parametri secondo il criterio della massima verosimiglianza.
Il metodo fornisce stime asintoticamente non distorte ed efficienti dei parametri, gestendo

contemporaneamente la presenza di dati mancanti e la stima dei parametri in un unico passaggio [4].

2.3 Standardizzazione dei dati

Dopo tutti questi passaggi, i dati vengono standardizzati secondo la formula:

$$Z = \frac{X - \mu}{\sigma}$$

Di seguito è mostrato l'output delle colonne standardizzate:

```
Total Deaths   No. Injured  No. Affected  Total Affected  \
count  2.726200e+04  2.726200e+04  2.726200e+04    2.726200e+04
mean  -2.085079e-18 -4.170158e-18  2.085079e-18    7.297776e-18
std    1.000018e+00  1.000018e+00  1.000018e+00    1.000018e+00
min   -5.247480e-01 -2.333599e+00 -2.635123e-01   -9.599076e-02
25%   -3.497277e-02 -5.270250e-02 -9.553214e-02   -9.593828e-02
50%   -3.461332e-02 -3.577826e-02 -9.170008e-02   -9.204515e-02
75%   -3.133341e-02 -3.149590e-02 -3.104052e-04   -5.065650e-05
max    8.308583e+01  7.722262e+01  6.561056e+01    6.500981e+01

Total Damage, Adjusted ('000 US$)  duration_days
count                       2.726200e+04   2.726200e+04
mean                        6.255236e-18   2.241460e-17
std                         1.000018e+00   1.000018e+00
min                        -1.710448e+00  -8.861385e-02
25%                        -5.146009e-02  -8.861385e-02
50%                        -1.387546e-02  -8.861385e-02
75%                        -2.649865e-03  -8.306026e-02
max                         1.033368e+02   1.419390e+02
```

2.4 Discretizzazione dei valori

Prima di procedere con le analisi delle regole di associazione, ho discretizzato i valori delle variabili quantitative: Total Deaths, No. Injured, No. Affected, Total Affected, Total Damage e duration_days considerando i tre quantili.
In pratica, ogni variabile è stata suddivisa in tre intervalli di uguale dimensione (basso, medio, alto) in base alla distribuzione dei dati, e quindi si tratta del metodo "Equal Depth". In questo modo, a ciascun intervallo è stata associata una valutazione della gravità dei danni o dell'impatto, classificata come Low, Med o High.

3 Clustering

Il clustering rappresenta uno step fondamentale nell'analisi e nella scoperta di pattern all'interno del database. Per questo motivo ho deciso di sperimentare 4 tecniche di clustering diverse: K-Means, Agglomerative Clustering, GMM e DBSCAN.
I parametri di questi algoritmi sono stati selezionati attraverso una procedura di fine tuning. Ho definito una griglia di possibili valori e utilizzando la metrica del Silhouette Score, è stata scelta la combinazione che produce il valore più alto.
Successivamente, con i parametri ottimali individuati, è stata condotta l'analisi. Per comprendere meglio la suddivisione dei dati, ho visualizzato i risultati in due dimensioni (scatterplot tramite PCA) e ho analizzato anche il coefficiente di silhouette.

3.1 K-Means

I parametri che ho sperimentato per il K-Means sono il numero di clusters e il metodo di inizializzazione. Per quanto riguarda il numero di clusters, ho provato con 2, 3 e 4, poiché un numero maggiore di 4, rende l'interpretazione dei risultati più complessa (o impossibile). Ho inoltre sperimentato se l'inizializzazione dovesse essere casuale oppure basata su K-Means++, che seleziona i centroidi iniziali utilizzando un campionamento più intelligente.

La combinazione restituita come quella migliore è:
`{'init': 'k-means++', 'n clusters': 3}`

3.1.1 Interpretazione dei cluster

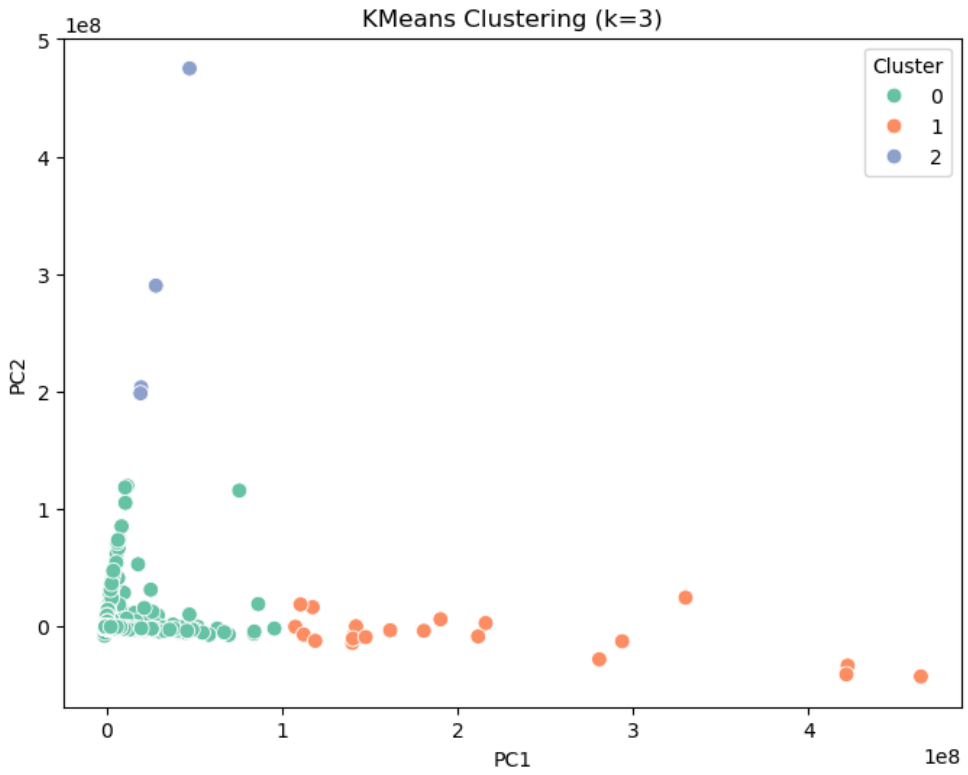

Figure 18: Kmeans clustering per n = 3

Da questo grafico, la suddivisione in tre cluster appare sensata anche visivamente. Il primo gruppo, o “Gruppo 0”, è rappresentato dalla concentrazione di punti vicino all’origine, mentre gli altri due gruppi corrispondono ai punti distribuiti a destra e a sinistra del gruppo principale. Sarebbe molto interessante analizzare le caratteristiche dei due gruppi che si discostano in questo modo da quello più numeroso.
Sul file Jupyter Notebook ho stampato le righe corrispondenti agli esempi presenti in ciascun cluster.
Nel cluster 1, rappresentato dai punti arancioni, si nota che tutti i disastri appartengono alla categoria "Naturali" e si concentrano principalmente in India e in Cina. Questi eventi sembrano avere in comune un numero molto elevato di danni economici e umani, oltre a una durata prolungata nel tempo.
Nel cluster 2, rappresentato dai punti blu, si trovano solo 4 esempi. I disastri appartenenti a questo gruppo rientrano nella categoria "Naturali" e si distinguono per l’elevato ammontare dei danni economici. Questi eventi si sono verificati in Giappone, negli Stati Uniti e in Perù, e ciò potrebbe spiegare la quantità dei danni economici, considerando le maggiori risorse economiche di Giappone e Stati Uniti. In particolare, si tratta di movimenti del suolo e tsunami per il Giappone, di cicloni tropicali per gli Stati Uniti, mentre il Perù è stato colpito da un’ondata di freddo. Tutti questi disastri appartengono ai sottogruppi dei disastri geofisici e meteorologici.

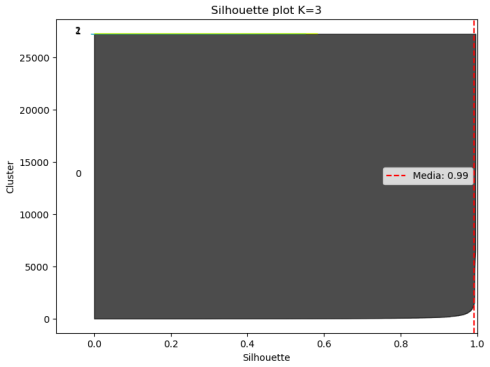

Figure 19: Silhouette plot k = 3

Considerando la maggiore concentrazione dei punti nel cluster 0 e la qualità della suddivisione dei cluster (Sil_score = 0.9925), il grafico del coefficiente di silhouette fornisce un’informazione limitata. Ciò è dovuto al forte squilibrio nella numerosità dei cluster. Il primo conta circa 25.000 elementi, mentre l’ultimo ne contiene soltanto 4.

3.2 Gaussian Mixture Models

Tramite l'utilizzo del clustering probabilistico di Gaussian Mixture Models, cerchiamo di rilassare il vincolo della forma sferica dei clusters e con varianza uguale in tutte le direzioni. Si assume che i dati provengono da una combinazione di distribuzioni gaussiane multivariate, utilizzando la massima verosimiglianza per stimare i paramteri.
La combinazione migliore degli iperparametri restituita per il GMM è `'covariance_type': 'tied', 'n_components': 2` con un silhouette score pari a 0.9922. L'algoritmo ha selezionato come migliore la covarianza di tipo "tied", il che significa che si assumono matrici di covarianza condivise per entrambi i cluster. Il numero ottimale di cluster individuato è pari a 2, inferiore rispetto a quanto ottenuto con K-means. Di seguito cercherò di interpretare questo risultato.

3.2.1 Interpretazione dei cluster

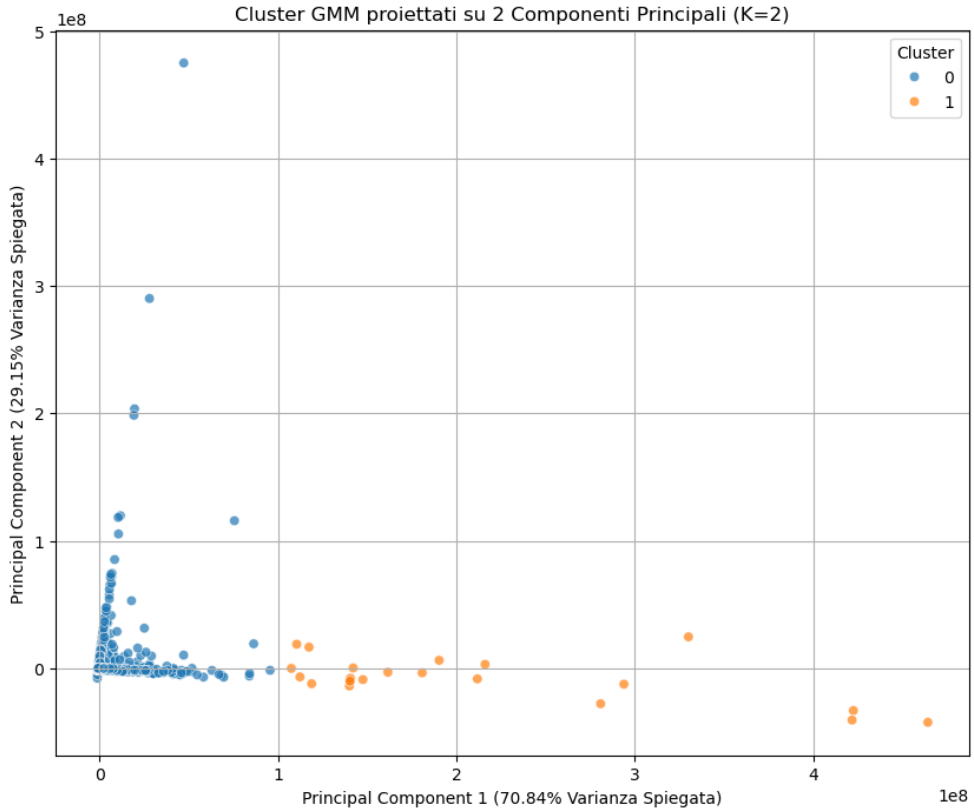

Figure 20: Cluster GMM proiettati

Possiamo osservare subito che il metodo dei Gaussian Mixture Models non è stato in grado di cogliere le sottili differenze tra il primo e il secondo gruppo, che K-means era riuscito a distinguere. In particolare, la suddivisione dei cluster individuata da GMM si basa sul principio della gravità complessiva dei disastri. Ciò significa che, a partire dal Gruppo 0, il più numeroso, è stato distinto solo il Gruppo 1, che contiene i disastri con i valori più elevati di danni umani ed economici e con una durata maggiore. Questi disastri sono concentrati nel territorio asiatico, in particolare in Cina e India, e si riferiscono a situazioni estreme di siccità e alluvioni.

3.3 Agglomerative Clustering

L'Agglomerative Clustering è una tecnica di clustering basata su un approccio gerarchico. I dati vengono raggruppati progressivamente partendo da singoli punti fino a formare cluster più grandi. I parametri che ho scelto di sperimentare includono il numero di cluster (con valori 2, 3 e 4), la metrica di distanza tra i punti, (considerando sia Euclidea che Manhattan), e il criterio di collegamento (linkage) tra i cluster, valutando le opzioni di complete, average e single.
Gli iperparametri ottimali, che ci hanno permesso di ottenere un valore di silhouette pari a 0.9961, sono:
`'linkage': 'average', 'metric': 'euclidean', 'n_clusters': 2`
Questo significa che la migliore divisione dei dati in due cluster è ottenuta utilizzando la distanza euclidea tra i punti, mentre il criterio di collegamento "average" permette di calcolare la distanza tra cluster come la media di tutte le distanze tra i punti dei due cluster.

3.3.1 Interpretazione dei cluster

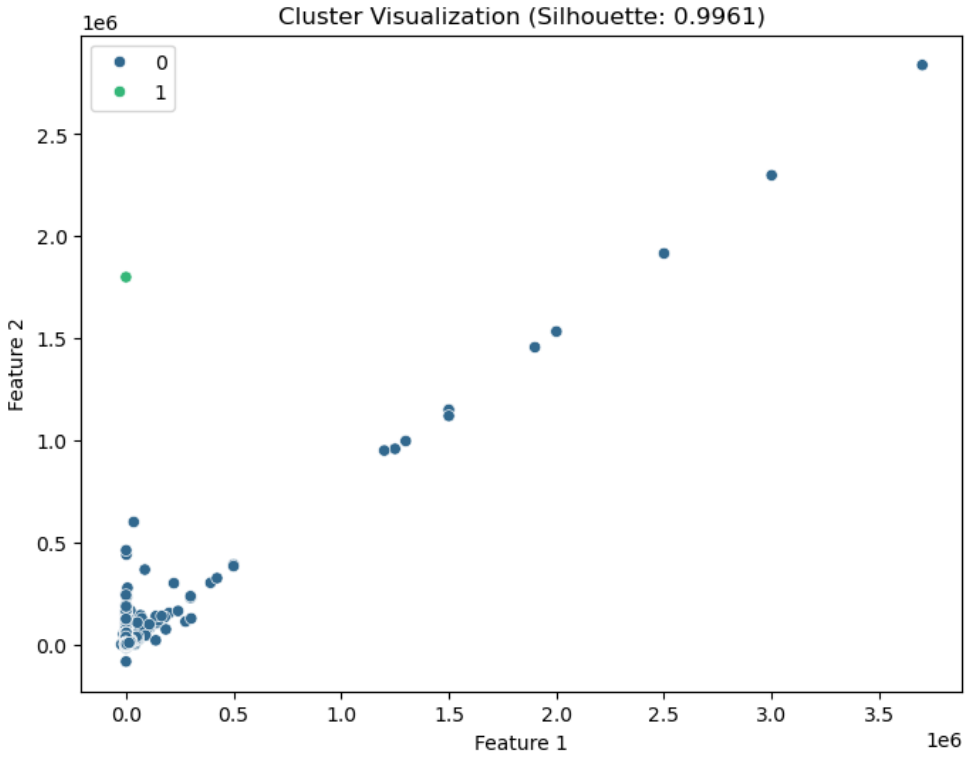

Figure 21: Visualizzazione di Agglomerative clustering

Come si può notare, la divisione ottenuta con questo tipo di clustering non è soddisfacente, perché la maggior parte dei dati è stata raggruppata in un unico cluster, a differenza di quanto ci si aspetterebbe.

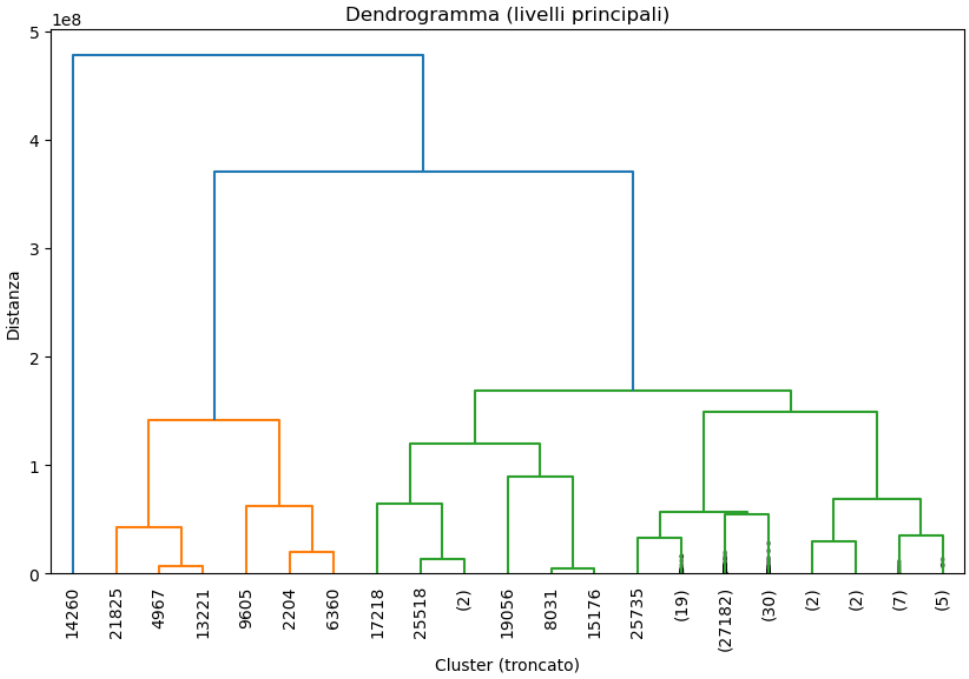

Figure 22: Livelli principali della dendrogramma

Anche dal dendrogramma si osserva che un punto non si collega agli altri cluster fino allo stado finale dell'algoritmo.
Questo punto rappresenta il disastro meteorologico verificatosi in Perù, colpito da freddo estremo. Il disastro ha causato 90 morti, 1.800.000 feriti e 337.467 persone direttamente colpite, con un danno economico totale di 478.711.300.000 US$ e una durata di 30 giorni.

3.4 DBSCAN

La tecnica di DBSCAN richiede due parametri: ϵ e il numero minimo di punti necessari per formare una regione densa (minPts). In questa analisi, ho sperimentato diversi valori di ϵ, differenti quantità minime di punti e le distanze Manhattan ed Euclidea.
Gli iperparametri ottimali individuati sono stati:
`'eps': 0.3, 'metric': 'euclidean', 'min_samples': 3`
, che hanno garantito un silhouette score (escludendo gli outlier) di 0.9998.

3.4.1 Interpretazione dei risultati

Questo algoritmo ha identificato 858 cluster e 16.931 punti considerati rumore.
Si nota che, a causa dell'elevato numero di cluster, visualizzarli non è pratico. Di conseguenza, questa tecnica non ha prodotto risultati soddisfacenti ai fini dell'analisi.
In conclusione, la tecnica migliore si è rivelata K-means, nonostante sia la più semplice, è stata in grado di cogliere le sottili differenze tra i vari disastri e di produrre tre cluster significativi.

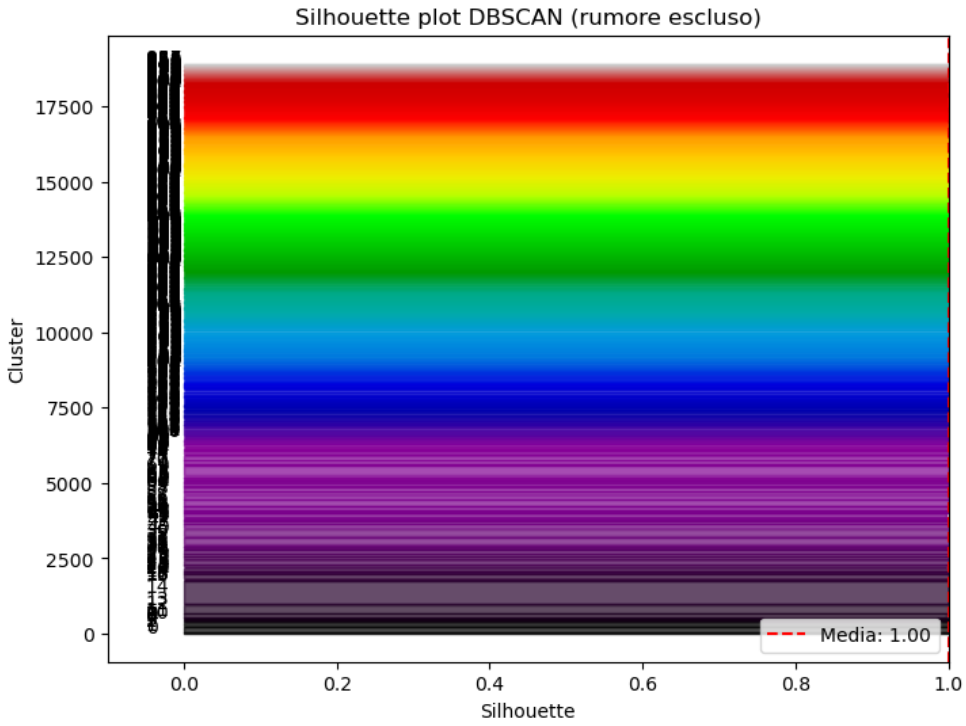

Figure 23: Silhoette score di DBSCAN

4 Regole di associazione

In questa sezione ci focalizziamo sulla scoperta dei pattern tramite algoritmi di Apriori e FP-growth. L'obiettivo e quello di scoprire nuovi pattern e regole di associazione all'interno del database.

4.1 Algoritmo Apriori

Apriori è un algoritmo per l'estrazione di set di elementi frequenti e l'apprendimento di regole di associazione su database relazionali. Procede identificando i singoli elementi frequenti nel database e estendendoli a set di elementi sempre più grandi, purché tali set compaiano con sufficiente frequenza nel database. I set di elementi frequenti determinati da Apriori possono essere utilizzati per individuare regole di associazione che evidenziano tendenze generali nel database.

4.1.1 Livelli di confidenza e supporto

Per migliorare la qualità delle regole, ho sperimentato con diversi livelli di confidenza e supporto.

```
support_levels = [0.03, 0.05, 0.07]
confidence_levels = [0.6, 0.7, 0.8]
```

Tra i primi risultati di questa tecnica, notiamo che alcune regole sono ridondanti e semplicemente evidenziano la struttura gerarchica del database.

```
SE {'Disaster Subtype_Fire (Miscellaneous)'}
ALLORA {'Disaster Type_Fire (Miscellaneous)',
        'Disaster Group_Technological'}
support=0.03, confidence=1.00, lift=32.73

SE {'Disaster Type_Fire (Miscellaneous)'}
ALLORA {'Disaster Subtype_Fire (Miscellaneous)',
        'Disaster Group_Technological'}
support=0.03, confidence=1.00, lift=32.73

SE {'Disaster Subtype_Fire (Miscellaneous)',
        'Disaster Group_Technological'}
ALLORA {'Disaster Type_Fire (Miscellaneous)',
        'Disaster Subgroup_Miscellaneous accident'}
support=0.03, confidence=1.00, lift=32.73...
```

Queste regole non costituiscono conoscenza nuova. Questo è un problema affrontato e risolto in seguito.
Tuttavia, tra le prime regole generate possiamo trovare anche dei pattern interessanti.

```
SE {'Total Affected_High', 'Disaster Type_Air',
        'No. Affected_High', 'Disaster Group_Technological'}
ALLORA {'Total Damage_High', 'Disaster Subtype_Air'}
support=0.03, confidence=1.00, lift=31.19
```

Si osserva un pattern dominante che collega i disastri di sottotipo "Air" a danni costantemente elevati, sia in termini economici sia nel numero di persone colpite. Questa regola presenta un supporto di 0.03 e una confidenza elevata di 1.00.
Per individuare strutture più significative, ho deciso di analizzare alcune colonne alla volta.

4.1.2 Pattern orientati

Per valutare meglio l'impatto dei disastri, l'analisi è stata ripetuta considerando soltanto le colonne: *Disaster Type_*, *Total Deaths_*, *Total Damage_*, *Total Affected_*.
Sperimentando nuovamente con diversi livelli di supporto e confidenza, ho scoperto un pattern rilevante:

```
SE {'Disaster Type_Road'}
ALLORA {'Total Damage_Low', 'Total Affected_Low'}
support = 0.07, confidence = 0.61, lift = 3.53

SE {'Total Damage_Low', 'Total Deaths_Med',
        'Disaster Type_Road'}
ALLORA {'Total Affected_Low'}
support = 0.04, confidence = 1.00, lift = 3.00
```

Da questa regola possiamo capire che gli incidenti stradali nella maggior parte dei casi, sono a basso impatto economico e colpiscono un numero limitato di persone.
Tuttavia, la letalità a livello medio (*Total Deaths_Med*) suggerisce che, anche se coinvolgono poche persone, gli incidenti stradali tendono ad avere un numero di morti non trascurabile rispetto all'impatto economico.
Inoltre, notiamo la presenza di diversi pattern simmetrici, ad esempio:

```
SE {'Total Affected_High', 'Disaster Type_Air'}
ALLORA {'Total Damage_High'}
support=0.03, confidence=1.00, lift=3.00

SE {'Total Affected_High', 'Disaster Type_Road'}
ALLORA {'Total Damage_High'}
support=0.04, confidence=1.00, lift=3.00

SE {'Disaster Type_Water', 'Total Affected_High'}
ALLORA {'Total Damage_High'}
support=0.04, confidence=1.00, lift=2.99
```

Questi pattern simmetrici suggeriscono che in vari tipi di disastro, se le persone colpite sono molte, anche i danni economici sono elevati. Quindi, la relazione tra "affected" e "damage" non dipende dal tipo di disastro, è robusta e generalizzabile.
Inoltre, in questa sezione cercherò di esplorare altri pattern riguardanti le regioni e i paesi in cui si verificano i disastri. Pertanto, considero queste colonne da inserire nell'algoritmo: *Disaster Type_*, *Subregion_*, *Total Deaths_*, *Total Damage_*.
Notiamo un pattern geografico molto forte:

```
SE {'Disaster Type_Epidemic'}
ALLORA {'Subregion_Sub-Saharan Africa'}
support=0.03, confidence=0.57, lift=3.21...
```

Oltre la metà delle epidemie nel database avviene in Africa Sub-Sahariana, e il lift = 3.21 indica che questa associazione è oltre tre volte più probabile della media.

```
SE {'Disaster Type_Air'}
ALLORA {'Total Damage_High'}
support=0.03, confidence=0.79, lift=2.38
```

Riappare il pattern dei disastri aerei, dove il 79% causa danni economici circa 2.4 volte più spesso del normale. Anche se i disastri aerei hanno bassa frequenza, presentano un'alta intensità economica.
Inoltre, notiamo pattern di gravità media, come:

```
SE {'Disaster Type_Epidemic'}
ALLORA {'Total Damage_Med'}
support=0.04, confidence=0.78, lift=2.33

SE {'Disaster Type_Flood'}
ALLORA {'Total Damage_Med'}
support=0.11, confidence=0.51, lift=1.52
```

Un'altra regole interessante:

```
SE {'Subregion_Northern America'}
ALLORA {'Total Damage_High'}
support=0.03, confidence=0.51, lift=1.52
```

Nella macroregione Nord America, i disastri tendono più spesso a essere costosi. Essendo paese ad alto reddito, questa regola riflette l'effetto del valore delle infrastrutture.
Un altro pattern interessante per i disastri del tipo "Water" è:

```
SE {'Disaster Type_Water'}
ALLORA {'Total Damage_High'}
support=0.04, confidence=0.68, lift=2.03

SE {'Disaster Type_Water'}
ALLORA {'Total Deaths_Med'}
support=0.04, confidence=0.60, lift=1.79
```

Questa regola riflette il fatto che i disastri acquatici comportano gravi danni economici, ma una letalità moderata.

4.2 Algoritmo FP-Growth

L'algoritmo FP-Growth (Frequent Pattern Growth) estrae in modo efficiente itemset frequenti da grandi database transazionali. A differenza dell'algoritmo Apriori, che presenta un elevato costo computazionale dovuto alla generazione di candidati e alle molteplici scansioni del database, FP-Growth evita queste inefficienze comprimendo i dati in un FP-Tree (Frequent Pattern Tree) ed estraendo i pattern direttamente da esso.
I pattern estratti da questo algoritmo non rappresentano conoscenze nuove, in quanto confermano le regole del algoritmo Apriori. Le regole ruotano attorno alla combinazione di disastri stradali con valori Low per danni, feriti e persone colpite.

4.2.1 Illustrazione dei pattern

```
SE {'Disaster Type_Road', 'No. Injured_Low'}
ALLORA {'Total Damage_Low', 'Total Affected_Low',
         'Disaster Subtype_Road'}
support=0.07, confidence=0.94, lift=14.00

SE {'No. Affected_Low', 'Disaster Type_Road',
         'No. Injured_Low', 'Disaster Group_Technological'}
ALLORA {'Total Damage_Low', 'Disaster Subgroup_Transport',
        'Total Affected_Low', 'Disaster Subtype_Road'}
support=0.07, confidence=0.94, lift=14.02

SE {'Total Affected_Low', 'Disaster Subtype_Road',
         'No. Injured_Low'}
 ALLORA {'No. Affected_Low', 'Total Damage_Low',
          'Disaster Type_Road'}
support=0.07, confidence=0.94, lift=14.02

SE {'Total Damage_Low', 'Total Affected_Low',
         'Disaster Type_Road', 'Disaster Subgroup_Transport'}
ALLORA {'No. Affected_Low', 'No. Injured_Low',
         'Disaster Subtype_Road',
         'Disaster Group_Technological'}
support=0.07, confidence=1.00, lift=14.02

SE {'No. Affected_Low', 'Total Damage_Low',
        'Disaster Subgroup_Transport', 'Disaster Type_Road'}
ALLORA {'Total Affected_Low', 'No. Injured_Low',
         'Disaster Subtype_Road',
         'Disaster Group_Technological'}
support=0.07, confidence=1.00, lift=14.02

SE {'Total Affected_Low', 'Disaster Type_Road',
         'No. Injured_Low'}
ALLORA {'No. Affected_Low', 'Total Damage_Low',
        'Disaster Subtype_Road'}
support=0.07, confidence=0.94, lift=14.02
```

```
SE {'No. Affected_Low', 'Disaster Type_Road',
         'No. Injured_Low'}
ALLORA {'Total Damage_Low', 'Total Affected_Low',
        'Disaster Subtype_Road'}
support=0.07, confidence=0.94, lift=14.02

SE {'Total Damage_Low', 'Total Affected_Low',
         'Disaster Subtype_Road'}
ALLORA {'No. Affected_Low', 'Disaster Type_Road',
         'No. Injured_Low'}
support=0.07, confidence=1.00, lift=14.02
```

References

[1] [Online]. Available: https://doc.emdat.be/docs/introduction/

[2] [Online]. Available: https://doc.emdat.be/docs/data-structure-and-content/disaster-classification-system/

[3] [Online]. Available: https://doc.emdat.be/docs/data-structure-and-content/emdat-public-table/

[4] [Online]. Available: https://files.emdat.be/docs/20240808%20Caveat%20of%20missing%20data%20in%20EM-DAT.pdf