MACHINE LEARNING E AIUTI ECONOMICI: UN'ANALISI TRAMITE INDICATORI NAZIONALI

Marco Braga*, Alessandro Maccario*, Oscar Zanotti*

*CdLM Data Science, Università degli Studi di Milano Bicocca

Abstract

Quali sono i paesi a livello mondiale con maggiore necessità di aiuti umanitari e finanziari? Cosa li distingue e li allontana dai paesi che presentano migliori condizioni economiche, sociali e sanitarie? Il seguente lavoro si è voluto incentrare sull'analisi di un dataset contenente svariati indicatori (sociali, economici, sanitari) relativi alla quasi totalità dei paesi esistenti, con lo scopo di individuare, tramite una cluster analysis, quali siano quei territori che maggiormente necessitano di soccorso (umanitario ed economico) proveniente, in special modo, dalle Nazioni Unite in collaborazione con "The World Bank Group", i quali hanno l'obiettivo di mantenere la pace e la sicurezza internazionale, migliorando le economie dei paesi che necessitano di risorse economiche e stabilendo decrescenti livelli di povertà. Tramite i risultati di questa tipologia di studio i *policy-makers* potrebbero più facilmente individuare le aree a livello mondiale che necessitano di maggiori aiuti finanziari in ottica di efficiente allocazione delle risorse economiche.

1 Introduzione

Dalla sua costituzione nel 1944 con gli accordi di Bretton Woods, "The World Bank Group" ha lavorato per aiutare più di cento stati in via di sviluppo permettendo loro di accedere a prestiti finanziari nei settori determinanti dell'economia, in grado di garantire una crescita economica tale da renderli autonomi. "The World Bank Group" lavora con governi nazionali, settore privato, organizzazioni della società civile, banche di sviluppo regionali, centri di studio e altre istituzioni internazionali su temi che spaziano dal cambiamento climatico, conflitti, sicurezza alimentare, educazione, agricoltura, finanza e scambi commerciali. Tutti questi sforzi supportano il duplice obiettivo del Bank Group di terminare entro il 2030 l'estrema povertà e aumentare la prosperità condivisa del 40% della popolazione più povero in tutte le nazioni.

Il dataset che si è deciso di analizzare comprende indicatori economici, sociali, ambientali e infrastrutturali della quasi totalità dei paesi a livello mondiale. L'obiettivo del successivo studio vuole essere quello, tramite una Cluster Analysis, di individuare quei soggetti che presentano una più alta necessità di aiuti economici tali che, l'eventuale aiuto da parte della World Bank possa essere mirato, continuo e più efficace tanto a livello statale che di settori specifici interni alla società stessa.

Posta la domanda "Cosa rende un paese povero tanto diverso da uno più ricco?", la successiva risposta vuole essere incentrata sull'individuazione di quelle caratteristiche che identificano uno stato con maggiore bisogno di aiuti economici ed umanitari, mirati al superamento della condizione di necessità degli individui.

Mentre le analisi della World Bank in termini di classificazione dei paesi a livello mondiale prendono in considerazione solamente il GNI (Gross National Income), la nostra ricerca vuole approfondire tale suddivisione per comprendere se non siano da considerare anche altri attributi necessari ad approfondire le dinamiche interne agli stati per essere in grado di indirizzare aiuti adeguati a coloro i quali necessitano in misura maggiore.

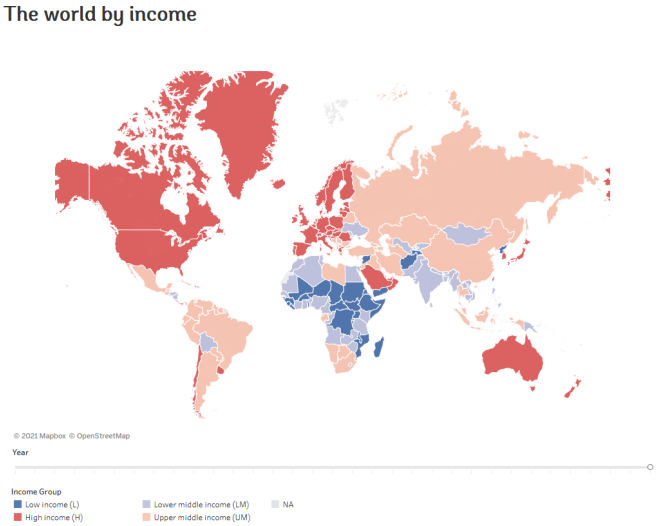

Figura 1: gross national income (GNI) per capita data in U.S. dollars

2 Descrizione data set

Attraverso la raccolta e l'elaborazione di dati provenienti da agenzie statistiche nazionali, banche centrali e custom services, "The World Bank Group" redige ogni anno "The World Development Indicators", una compilazione di statistiche rilevanti, di alta qualità e comparabili a livello internazionale sullo sviluppo globale e la lotta alla povertà, contenente indicatori sulla salute e l'educazione fino a informazioni relative all'economia. Nello specifico, il database contiene 1.400 indicatori di serie temporali per 217 economie e più di 40 gruppi di paesi, spesso con dati che risalgono a più di 50 anni fa.

L'anno di riferimento scelto per le analisi è il 2016 in quanto unico anno a contenere la percentuale minore di valori nulli, rispetto agli anni precedenti e successivi, ed essendo temporalmente prossimo alla data di analisi stessa.

3 Ristrutturazione data set

Il dataset originale consta di 64 colonne e 380161 osservazioni. Si è svolto un lavoro di eliminazione di quegli attributi e istanze non utili ai fini dell'analisi (i valori colonnari facevano riferimento tanto ai nomi dei paesi che agli indicatori ma anche a tutti gli anni di riferimento dei dati). Le prime 67681 righe contenevano dati inerenti le macro aree mondiali come, ad esempio, il mondo arabo o l'Unione Europea, ed è stata necessaria un'operazione di pivoting per ruotare la tabella originaria in quanto nella terza colonna, per ogni riga, erano presenti gli attributi suddivisi per singolo stato con i rispettivi valori degli indicatori nell'ultimo attributo.

Dopo una prima esplorazione dei dati, si è creato un nuovo data set contenente 35 stati appartenenti a due tipologie:

- Territori dipendenti: stati che non sono considerabili come indipendenti dalla madrepatria e che, quindi, mancavano di molti dei valori degli attributi facenti riferimento al paese di dipendenza;
- Microstati: in quanto caratterizzati da una struttura fiscale particolarmente favorevole, quindi non necessitanti di aiuti economici e mancanti un numero elevato di valori non facilmente reperibili.

Considerata l'importanza del valore del Prodotto Interno Lordo del paese nell'analisi per i motivi di erogazioni di aiuti economici ed umanitari, e poiché si presentavano mancanti solamente 6 valori, si è deciso di integrare tali elementi nel dataset originale tramite fonti ufficiali o attendibili in maniera tale da ridurre al minimo eventuali perdite importanti di informazione.

Inoltre, in questa fase di pre-processing sono stati individuati ulteriori altri dati utili e necessari all'elaborazione successiva, ovvero:

- crescita del PIL in percentuale;[1]
- PIL corrente;[1]
- Area territoriale (esclusi corsi d'acqua);[2]
- Area di superficie (inclusi i corsi d'acqua);[2]
- Popolazione totale;[3]
- Crescita della popolazione in percentuale;[3]
- Popolazione urbana dell'Eritrea;[3]
- Popolazione rurale dell'Eritrea;[3]
- Emissioni di CO2;[5]
- Percentuale di donne in parlamento.[4]

Per alcuni specifici attributi (fixed telephone subscription (% sulla popolazione), mobile cellular subscription, population 0-14 - total, population ages 65 and above, total) è stato deciso di trasformali da frequenze assolute a frequenze percentuali rispetto alla popolazione totale di ogni singolo stato. Nel caso del number of deaths ages 0-14 (% of pop) abbiamo invece creato una nuova colonna che riunisce tutte le info che contiene le % di numero di morti tra 0-14 sul numero della popolazione relativi a questa fascia d'età.

4 Selezione delle variabili

Tramite l'uso della piattaforma Knime (il software usato per lo svolgimento della seguente analisi) si sono scelti, dagli oltre 1.400 indicatori, quelli che presentavano una soglia di valori mancanti pari al massimo all' 1.5% del totale delle istanze presenti in ogni attributo; così facendo, si sono ricavate 166 variabili. Successivamente, data l'esistenza di variabili che sono state considerate dipendenti, è

stata approntata una cernita tramite la scelta degli attributi più interessanti giungendo ad un totale di 32 attributi totali e 180 paesi da valutare, dati considerati sufficienti per applicare le successive analisi di cluster.

5 Clustering

Prima di procedere con l'analisi si normalizzano gli attributi eseguendo una trasformazione lineare tramite il metodo Z-score. Al fine di ridurre il numero eccessivo di attributi correlati, è stato deciso di usare un correlation filter con il quale si sono esclusi gli attributi con una correlazione maggiore di 0.9. Si procede quindi con la Cluster Analysis sottoponendo il dataset, ripulito e normalizzato, a due diversi metodi di aggregazione: gerarchico, K-means.

5.1 Clustering gerarchico

I clustering gerarchici possono presentarsi sotto due forme: **agglomerativi**, dove ogni osservazione viene considerata come un cluster singolo per poi, tramite una procedura a step, raggruppare le istanze fino a costituire un unico cluster contenente tutte le osservazioni; **divisivi**, dove le osservazioni formano un unico cluster il quale, negli step successivi, viene suddiviso in cluster più piccoli.

Si è deciso di approntare un clustering di tipo agglomerativo in quanto più comunemente utilizzato in letteratura.

Per calcolare la prossimità tra i cluster è stato utilizzato il metodo di *Ward*, che misura la prossimità tra due cluster come l'incremento della somma degli scarti quadratici che risulta dalla fusione di due cluster. Per fare ciò, l'algoritmo necessita come input la matrice delle distanze $\mathbf{D} = \text{dist}(\mathbf{X})$, calcolata sulla base della distanza Euclidea.

Per la scelta del numero di cluster si è osservato il dendogramma in figura che mostra le relazioni tra cluster e sotto-cluster e l'ordine secondo i quale questi vengono aggregati.

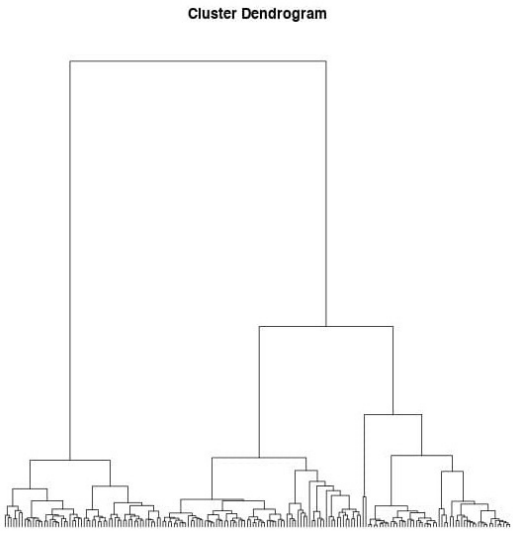

Figura 2: Dendrogramma del Clustering Gerarchico

Ogni ramo del dendogramma corrisponde a un possibile cluster. Sull'asse delle ordinate viene riportato il livello di distanza tra i vari gruppi, mentre sull'asse delle ascisse vengono riportati i singoli record. Tramite l'utilizzo di codice in linguaggio R è stato possibile quindi ottenere un taglio dell'albero di clustering all'altezza desiderata per evidenziare esattamente 4 cluster che verranno successivamente valutati.

5.2 K-means

L'approccio ai **Prototype-Based Clustering** si basa sull'assunzione che ogni cluster possa essere ben rappresentato da un unico punto chiamato **prototipo**. Ogni oggetto è quindi collocato nel cluster del prototipo a cui è più vicino. Il K-Means è un **algoritmo di apprendimento non supervisionato** in cui il prototipo prende il nome di *centroide*: questo valore solitamente è identificato dal vettore media dei valori degli attributi delle osservazioni di quel determinato cluster. L'algoritmo k-means è un algoritmo iterativo formato dai seguenti step:

1. **Inizializzazione**: occorre inizializzare l'algoritmo il numero iniziale dei **K** centroidi disposti casualmente, ovvero si scelgono il numero di cluster di cui il data set sarà composto;

2. **Assegnazione al Cluster**: l'algoritmo analizza ciascuno dei data points e li assegna al centroide più vicino. Viene quindi calcolata la distanza euclidea tra ogni record e ogni centroide. Ogni record sarà poi assegnato al centroide la cui distanza risulta minima;

3. **Aggiornamento della posizione del centroide**: nel caso in cui si siano formati nuovi cluster dal passaggio precedente, si ricalcola la posizione dei nuovi centroidi. Il nuovo valore di un centroide sarà la media di tutti i record che sono stati assegnati al nuovo cluster. L'algoritmo sarà iterato fino a che i centroidi non verranno nuovamente modificati, ossia quando si raggiunge un punto di convergenza tale per cui non si hanno più modifiche dei cluster.

Stabilire il numero di cluster iniziali è una parte fondamentale per questo algoritmo. Per compiere questa scelta ci si può affidare al risultato del cluster gerarchico, oppure si può fare affidamento sul coefficiente di Silhouette o l'indice di Dunn come mostrato di seguito:

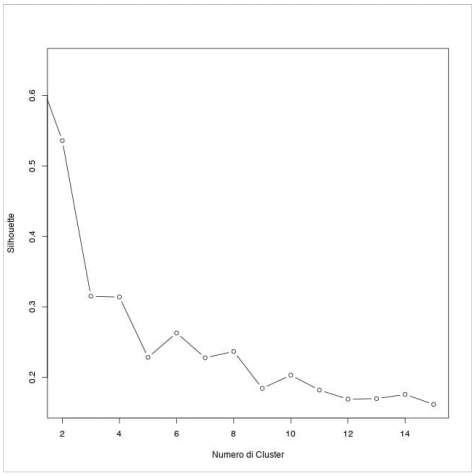

Figura 3: Indice di Silhouette rispetto all'algoritmo k-medie

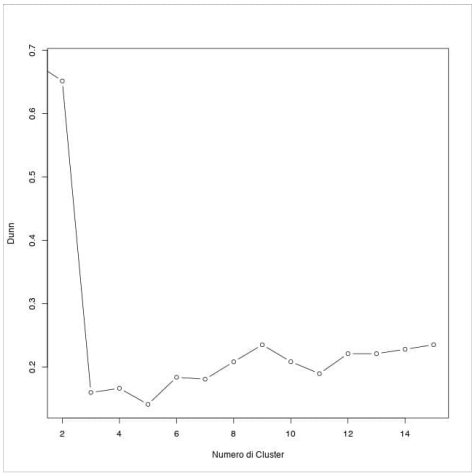

Figura 4: Indice di Dunn rispetto all'algoritmo k-medie

Entrambi gli indici devono essere massimizzati e quindi, come si osserva dalla figura [], il numero di cluster ottimale sembra essere 4. Infatti con 3 o 4 cluster si ottiene il massimo della Silhouette, escludendo i soli due cluster. Analizzando l'indice di Dunn, si può notare come la variazione dell'indice sia mimima oltre i quattro cluster. Così come per Ward, con l'obiettivo di utilizzare il maggior numero di informazioni disponibili, si suddivide il dataset in 4 cluster.

6 Validazione

Sebbene la cluster analysis, rispetto alle metodologie di classificazione, non abbia una fase di validazione così definita, si tratta di un procedimento di primaria importanza da attuare. Infatti ciascun algoritmo che ha come fine l'individuazione dei cluster, separa le osservazioni anche nel caso in cui non vi sia alcuna struttura nei dati; pertanto, è necessario valutare la bontà dei risultati. A questo scopo si sono considerate due tipologie di indicatori: esterni (supervisionati) e interni (non supervionati).

6.1 Indici esterni o supervisionati

Per poter mettere a confronto l'output dei due algoritmi scelti, e allo stesso tempo verificare che all'interno dei quattro clusters siano raggruppate veramente nazioni con un simile livello di sviluppo, si considera l'Indice di Sviluppo Umano, calcolato ogni anno dall'organizzazione delle Nazioni Unite[8], relativo all'anno 2016. Discretizzando la colonna "HDI" in 4 intervalli di pari ampiezza si è ottenuta la colonna "category". Considerando la partizione così ottenuta si possono suddividere le nazioni in 4 gruppi: *Indice basso, Indice medio, Indice alto, Indice molto alto.* Analizzando la partizioni e le suddivisioni nei cluster ottenute precedentemente, è possibile distinguere quattro situazioni per ciascuna coppia di osservazioni x e y:

(a) **x** e **y** appartengono allo stesso cluster e allo stessa partizione;

(b) **x** e **y** appartengono allo stesso cluster ma non alla stessa partizione;

(c) **x** e **y** appartengono alla stessa partizione ma non allo stesso cluster;

(d) **x** e **y** appartengono a cluster e partizione differenti.

1. Rand, definito come:

$$R = \frac{a+d}{a+b+c+d} \tag{1}$$

2. Jaccard, definito come:

$$J = \frac{a}{a+b+c} \tag{2}$$

3. Fowlkes and Mallows, definito come:

$$FM = \sqrt{\frac{a}{a+c} \times \frac{a}{a+b}} \tag{3}$$

Tabella 1: Indici esterni o supervisionati

Indici	**R**	**FM**	**J**
Ward	0.738	0.561	0.388
K-medie	0.741	0.574	0.4

I risultati ottenuti sono riportati nella Tabella 1. Si osserva che gli algoritmi sono caratterizzati da indicatori con valori alti e simili tra di loro. Poiché non è possibile scegliere l'algoritmo migliore solo sulla base di questi indici, si prosegue prendendo in considerazione gli indici interni o non-supervisionati.

6.2 Indici interni o non-supervisionati

Gli indici interni generalmente valutano la correttezza dei cluster tramite misure di *Coesione*, ovvero di quanto sono simili le istanze appartenenti allo stesso cluster, oppure di *Separazione*, ovvero quanto istanze appartenenti a cluster differenti sono distanti.

E' possibile quindi esprimere la *overall cluster validity* per un insieme di cluster come la somma pesata della misura di validità dei singoli cluster che può essere ottenuta, come nel caso del coefficiente di *Silhouette*, come una combinazione di queste due misure.

Nello specifico, il *Coefficiente di Silhouette* è così definito:

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)}[-1, 1] \quad (4)$$

dove:

- a_i è la distanza media della i-*esima* istanza da tutte le altre istanze presenti nel suo stesso cluster (come misura di *coesione*);
- b_i rappresenta il valore minimo delle distanze medie della i-*esima* istanza da tutte le altre istanze contenute in ogni cluster, eccetto quello che contiene la i-*esima* instanza stessa.

Oltre al coefficiente di Silhouette è stato calcolato il coefficiente di *Dunn* che assume valori nell'intervallo [0, $+\infty$] e definita come il rapporto tra la minore distanza tra le osservazioni che non si trovano nello stesso cluster e la massima distanza intra-cluster.[7]

I risultati vengono quindi riportati in tabella 2:

Tabella 2: Indici interni o non supervisionati

Indici	**Silhouette**	**Dunn**
Ward	0.213	0.133
K-medie	0.224	0.113

Per entrambi i coefficienti, valori alti indicano un buon funzionamento degli algoritmi utilizzati.

Nonostante le differenze siano minime, si rileva che il metodo Ward manifesti prestazioni inferiori all'algoritmo K-Medie sui dati in esame, per cui si sceglie quest ultimo output per il proseguo delle valutazioni.

6.3 Test formale

Di necessaria valutazione è l'esistenza o meno di una effettiva struttura all'interno del data set in studio, ovvero testare il *Validity Paradigm*. Nel caso non fosse possibile affermarlo, l'intera analisi di cluster non avrebbe sortito nessun guadagno. Per fare ciò, è stato quindi fondamentale eseguire un test in cui l'ipotesi nulla H0 consistesse nella *Random Position Hypothesis*, ovvero un test che ha il compito di controllare che la posizione delle **m** istanze in una specifica regione di uno spazio n-dimensionale siano equiprobabili.

Viene quindi sfruttato il coefficiente di Silhouette dell'algoritmo gerarchico per testarlo rispetto all'ipotesi nulla. A tal scopo, il metodo di Monte Carlo che genera una distribuzione empirica su 1000 simulazioni, permette il confronto del quantile di tale distribuzione (a livello $\alpha = 0.01$) con il valore della statistica-test, ovvero il coefficiente di Silhouette.

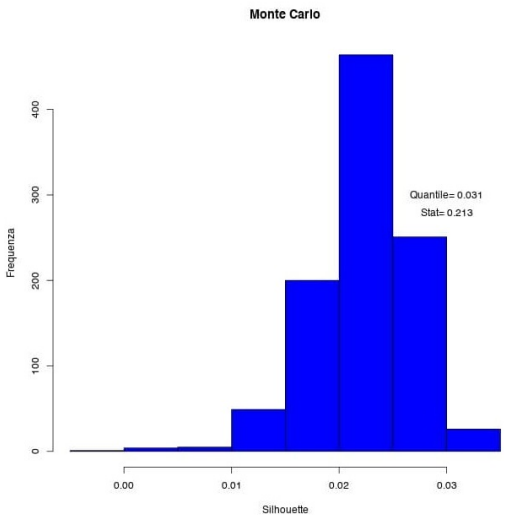

Figura 5: Distribuzione empirica dell'indice di Silhouette ottenuta con il metodo Monte Carlo

Essendo il quantile, pari a 0.031, inferiore al valore del coefficiente di Silhouette, pari a 0.213, viene rifiutata l'ipotesi nulla H0 di assenza di struttura all'interno dei dati, potendo affermare che quanto svolto nel processo di clusterizzazione abbia effettivamente avuto significato.

7 Interpretazione dei risultati

In seguito all'analisi di clustering si sono individuati quattro gruppi in cui sono suddivisi gli stati, ognuno dei quali con le proprie caratteristiche. In questa sezione del report si vogliono indagare alcune relazioni che intercorrono tra i quattro cluster e alcune delle variabili di interesse del dataset.

Avendo usato le k-medie abbiamo ottenuto le medie di ogni cluster per attributo e ottenuto quanto segue:

Cluster	**E**	**L**	**DW**	**P**
1	0.564	0.352	0.543	-0.04
2	-0.375	-0.599	-0.507	0.4
3	0.656	0.949	0.686	-0.866
4	-1.774	-1.589	-1.63	1.067

dove:

- E = electricity
- L = life expectancy
- DW = Drinking water
- P = population growth

Nello specifico, gli elementi del *cluster 3* si individuano gli stati maggiormente sviluppati in quanto presentano una media di valori associati alla tabella più alti, esclusa la crescita della popolazione che, in questi paesi, è molto bassa (ad esempio: Italia, Cina, Regno Unito). In tale cluster sono presenti sia le grandi potenze mondiali che presentano un GDP notoriamente elevato, come la Cina o

gli Stati Uniti con il GDP più elevato tra tutte le nazioni, sia stati come il Montenegro che, al contrario, presenta un basso GDP ma un'indicatore dell'accesso all'elettricità decisamente elevato o anche l'accesso all'acqua potabile. Allo stesso modo, un numero considerevole di paesi appartenente all'Eurozona sono presenti al suo interno.

Riguardo il *cluster 1*, si trovano paesi come il Brasile con elevato GDP, ma più bassi valori dell'accesso all'elettricità o alla disponibilità di acqua potabile. Al contrario, vi sono presenti paesi come il Kuwait con elevato indice di acqua potabile ed elettricità ma al contempo un GDP relativamente più basso.

Nel caso del *cluster 2*, dove vi sono paesi come il Rwanda presentano accesso all'elettricità e acqua, ma anche il GDP decisamente inferiore. Oppure l'India con un GDP molto elevato ma un accesso ai servizi come acqua ed elettricità; paesi che avrebbero bisogno di aiuti specifici.

Infine nel *cluster 4* si trovano paesi africani caratterizzati oltre che da un basso GDP anche da una mancanza nell'accesso ai servizi essenziali e caratterizzati da un elevata crescita della popolazione, unico attributo maggiore per i paesi più poveri e minore per i paesi più sviluppati.

Di conseguenza, gli aiuti dovrebbero essere indirizzati in prevalenza nel cluster 4 senza tralasciare aiuti specific per gli stati appartenenti al cluster 2.

8 Principal Component Analysis

Un ulteriore approccio che si è voluto impiegare per tentare di analizzare il fenomeno più approfonditamente è l'applicazione della procedura denominata *Principal Component Analysis*[6]. Le componenti principali che si ottengono dopo l'applicazione del metodo rappresentano combinazioni lineari della matrice X di partenza incorrelate fra di loro e in grado, una volta riordinate, di racchiudere al loro interno la massima varianza spiegabile da ogni singola componente. Si ottiene quindi una riduzione di dimensionalità del data set originale, guadagnando in termini di interpretabilità finale dell'analisi.

Poiché le componenti principali sono in numero pari agli attributi iniziali scelti, resta da valutare il numero di PC da mantenere per il proseguo della ricerca. Si è scelto quindi di utilizzare lo strumento della varianza spiegata, ovvero un *line plot* e un *bar plot* rappresentante la varianza spiegata da ogni componente.

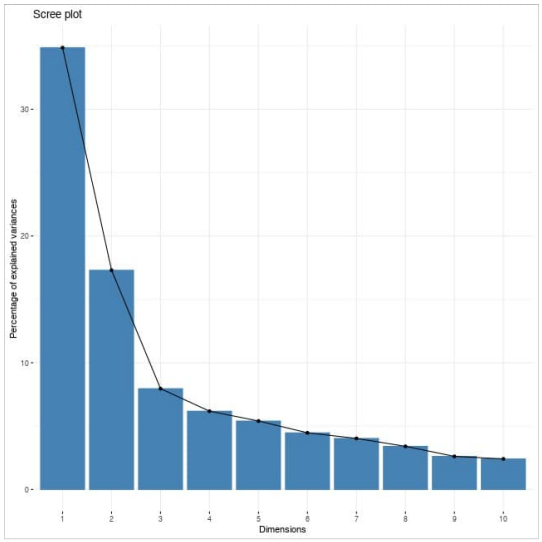

Figura 6: Grafico della varianza spiegata

La variazione maggiore di pendenza del grafico permette di scegliere il numero di componenti principali da mantenere pari a 3, a cui corrisponde il valore percentuale di varianza spiegata maggiore.

Dai dati risulta che la prima componente è correlata negativamente con "Life expectancy at birth", "Individual using the Internet", "People using at least basic drinking water services", "Access to electricity", e positivamente con "Adolescent fertility rate", "Prevalence of anemia among children" e "Age dependency ratio, young". La seconda PC invece dipende dagli attributi "Population ages 65 and above", "Population ages 0-14", "Fixed telephone subscriptions", "CO2 Emissions", "Population total". La terza PC a sua volta ha le correlazioni più forti con "Death rate" e "PM2.5 air pollution".

Attributi	**PC 1**	**PC 2**	**PC 3**
DW	-0.86	0.12	-0.27
LE	-0.92	0.11	-0.17
P	-0.09	-0.84	-0.13
INT	-0.88	0.11	-0.06
CO2	-0.25	-0.93	-0.04
AN	0.91	-0.08	0.01
AGE	0.93	-0.09	0.08
GDP	-0.32	-0.81	0.09
DR	-0.04	-0.03	0.81

dove:

- P = population total
- LE = life expectancy
- DW = Drinking water
- INT = individual using internet
- CO2 = CO2 emissions
- AN = Anemia among children

- AGE = Age dependency
- GDP = GDP current
- DR = Deaths rate

Una prima possibile interpretazione potrebbe quindi essere la seguente: la **prima componente** ingloba al suo interno tutti gli attributi socio-economici; la **seconda componente** mette in rilievo la distribuzione anagrafica della popolazione; la **terza componente**, infine, oltre ad aggregare i fattori già detti, tiene conto del tasso di morte e dell'inquinamento dell'area, che sono ovviamente correlati.

8.1 Analisi di clustering sulle Principal Components

A questo punto è possibile effettuare clustering sui punteggi delle prime tre componenti. L'algoritmo scelto è di tipo gerarchico, in particolare si utilizza il metodo di Ward. Dal dendrogramma emergono quattro clusters ben definiti.

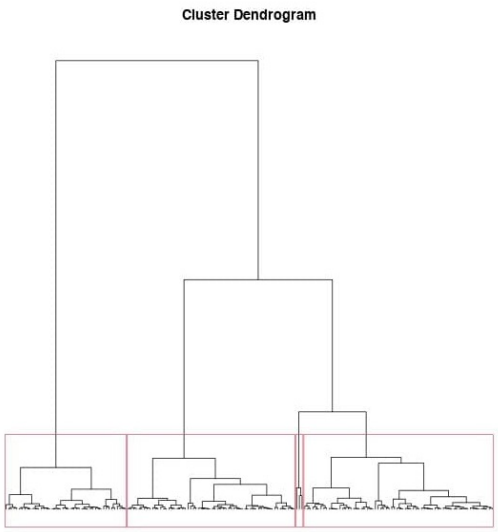

Figura 7: Dendrogramma della Principal Component Analysis

E' possibile sottolineare alcune differenze nella formazione dei cluster.

I risultati del raggruppamento con il metodo di Ward forniscono molti spunti interessanti, raggiunti solo grazie all'output di un'Analisi delle componenti principali. Si nota la presenza di un cluster con molti meno stati rispetto al clustering effettuato senza considerare il metodo della PCA: tramite questa procedura si identificano le super potenze mondiali (India, Cina, Usa) che altrimenti sarebbero indicate solo come stati sviluppati. Allo stesso modo, viene meno il cluster degli stati a rischio identificando in misura migliore gli stati ampiamente sviluppati da quelli che, effettivamente, necessitano di aiuti economici e umanitari.

9 Criticità e limiti

Nonostante l'individuazione corretta di molti cluster, l'analisi non è esente da criticità e limiti. Dal data set originale erano molti i valori mancanti rispetto, soprattutto, agli stati più piccoli (come i microstati) in settori quali l'economia, il lavoro, il servizio sanitario, quello educativo e quello ambientale: avendo a disposizione tali elementi si sarebbero potuti raggiungere differenti conclusioni permettendo un aiuto più diretto verso quelle istanze che necessitano di aiuti più mirati.

10 Conclusioni

La cluster analysis ha suddiviso gli stati in quattro gruppi con composizioni molto diverse: differiscono tra di loro per i valori economici, per l'accesso ai servizi essenziali, la crescita della popolazione o fattori ambientali. Queste informazioni potrebbero essere impiegate, come auspicato nell'introduzione, dalla "World Bank" o dalle Nazioni Unite al fine di adottare misure in termini di aiuti ai paesi identificati come i maggiormente sottosviluppati, o come a rischio, con l'obiettivo di migliorare. L'analisi delle componenti principali ha permesso di suddividere adeguatamente i diversi cluster permettendo una migliore analisi

E' emerso che gli stati maggiormente a rischio sono lacunosi dal punto di vista dei servizi essenziali come quelli sanitari e dei beni di prima necessità (come l'acqua potabile).

Riferimenti bibliografici

[1] Country Economy `https://countryeconomy.com/gdp`

[2] Surface and land `https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_area`

[3] Population `https://www.worldometers.info/world-population/population-by-country/`

[4] Women in parliament `https://gecpdsomalia.org/moving-towards-30-women-win-big-in-\somalias-2016-elections/`

[5] CO2 Emission `https://ourworldindata.org/co2-emissions`

[6] Principal Component Analysis `http://pzs.dstu.dp.ua/DataMining/pca/bibl/Principal%20components%20analysis.pdf`

[7] clValid an R package for cluster validation `http://lib.stat.cmu.edu/R/CRAN/web/packages/clValid/vignettes/clValid.pdf`

[8] Human Development Reports `http://hdr.undp.org/en/indicators/137506#`