LIFE EXPECTANCY

Analisi dell'impatto delle variabili economiche, sanitarie e demografiche sull'aspettativa di vita dei paesi del mondo.

Matteo Paolo Dell'Acqua, Mat. 875152
Cristian Mazzina, Mat. 864687
Karyna Petrashchuk, Mat. 864728

1. INTRODUZIONE

Il dataset oggetto della nostra analisi comprende dati relativi a variabili di natura economica, sanitaria, demografica e sociale che potrebbero avere un impatto sull'aspettativa di vita media di un paese che è un indicatore sintetico che riflette lo stato di salute e di sviluppo di una popolazione. I dati sono stati raccolti nel periodo 2000-2015 per 193 stati del mondo, con l'esclusione di alcuni paesi per i quali non erano disponibili informazioni sufficienti o affidabili (es. Sudan e Corea del Nord).

L'obiettivo della nostra analisi è quello di classificare i paesi in fasce di aspettativa di vita (<60, 60-70,70-78,>78) e capire se ci sono variabili non scontate che caratterizzano le varie fasce, la classificazione verrà effettuata sui dati relativi al 2015. Successivamente andremo a compiere una clusterizzazione non supervisionata per vedere se l'algoritmo capta dei cluster coerenti con la nostra classificazione.

La nostra analisi può avere un **significato** sociale ed economico importante, in quanto può fornire una maggiore comprensione dei fattori che influenzano l'aspettativa di vita dei paesi e delle disparità esistenti tra le diverse fasce. Questo può aiutare a identificare le priorità e le sfide per migliorare la salute e il benessere delle popolazioni, soprattutto quelle più vulnerabili e svantaggiate. Inoltre, può contribuire a sensibilizzare l'opinione pubblica e le istituzioni sul tema dell'aspettativa di vita.

2. MATERIALE E METODI

I dati sono basati sulle stime globali della salute dell'Organizzazione Mondiale della Sanità (OMS), che tiene traccia dello stato di salute e di molti altri fattori correlati per tutti i paesi. Il dataset è stato scaricato da Kaggle.com è composto da 21 variabili e 2864 osservazioni, per vedere i dettagli sulle variabili andare al link in appendice.

I metodi che abbiamo utilizzato sono model-based classification attraverso una mixture discrimant analysis, per fare ciò abbiamo utilizzato sia gli EDDA models che gli MDA models. Per far sì che tutti gli stati presenti nel dataset vengano classificati abbiamo creato un test set con i dati relativi al 2015, l'ultimo anno disponibile, ed usato tutte le altre osservazioni come training set. Dopodiché abbiamo applicato un model-based clustering per verificare se anche con questo metodo si riuscisse a cogliere la suddivisione da noi ipotizzata della variabile *Life expectancy.* Per fare ciò abbiamo aggiunto una colonna al nostro dataset riqualificando la variabile *Life expectancy* nelle seguenti classi: <60, 60-70,70-78,>78.

3. CONSIDERAZIONI SULLE VARIABILI E ANALISI ESPLORATIVA

Data la numerosità delle variabili vogliamo capire quali sono più rilevanti ai fini della nostra analisi, l'obiettivo è semplificare la stima dei modelli per migliorarne le prestazioni. Le variabili con alta

correlazione ci hanno portato ad approfondire la loro natura, di conseguenza abbiamo escluso dall'analisi le variabili con misurazioni relative a fenomeni simili (informazioni ridondanti). Per esempio le vaccinazioni contro la poliomielite (Polio) e contro il tetano (Diphtheria) sono correlate al 95% e presentano correlazioni quasi uguali con altre variabili del dataset.

Variabili ridondanti: *Polio*, *Under five deaths*, *Infant deaths*, *Adult mortality*, *Thinness ten nineteen years*.

Inoltre, abbiamo deciso di eliminare dalla nostra analisi la variabile *Population mln* siccome non è un'informazione che può discriminare l'aspettativa di vita tra una nazione ed un'altra (per es. sarebbe utile avere densità di popolazione).

L'ultima modifica che abbiamo apportato al dataset è stata rimuovere le due variabili dicotomiche complementari *Economy status developing* e *Economy status developed.*

Nelle prime prove di stima dei modelli abbiamo notato che le la presenza di esse pregiudicava le nostre analisi, infatti qualsiasi classificazione che non fosse stata in due gruppi produceva un error rate significativamente più alto rispetto alla stessa classificazione fatta con esclusione di queste variabili, questo andava in contro all'obiettivo della nostra analisi. I metodi di classificazione e clustering sono stati implementati sulle variabili rimanenti

Dall'**analisi univariata** emerge che: dall'analisi dei boxplot si evince che tutte le variabili, tranne *Schooling*, hanno presenza di outlier, tuttavia essendo un numero molto alto decidiamo di non toglierli, siccome non crediamo siano derivati da errori d'inserimento, ma bensì dal significato della variabile.

Infatti in variabili come *GDP_per_capita* (*Fig. 1*) gli outlier sono dovuti ai pochi paesi nella quale c'è la ricchezza più alta.
Analizzando gli istogrammi delle variabili numeriche (*Fig. 2*) notiamo che alcune di esse sembrano presentare multimodalità, il che potrebbe farci pensare che i dati provengano da una mistura di più distribuzioni.

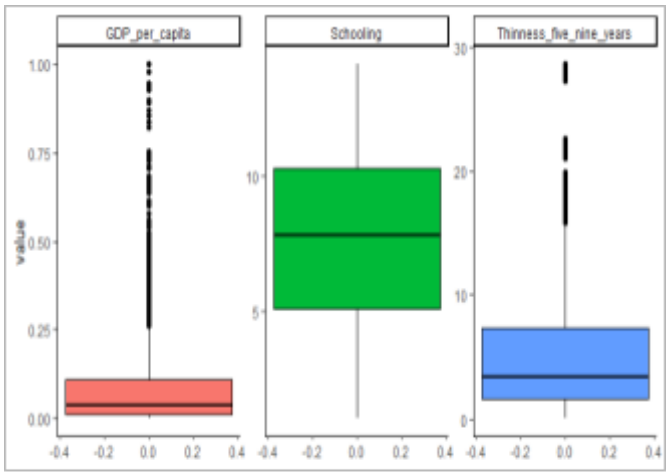

Fig. 1: Boxplot per le variabili GDP per capita, Schooling (anni di scuola) e Thinness 5-9 years (numero di bambini eccessivamente magri)

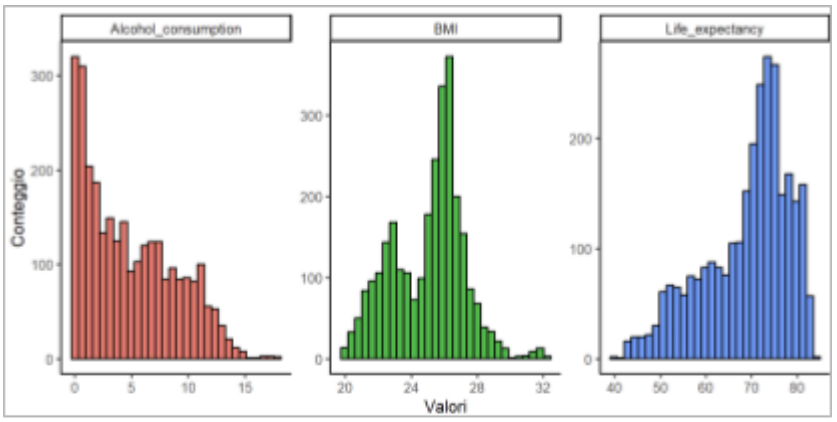

Fig. 2: istogrammi per le variabili *Alcohol consumption* (consumo di alcol), *BMI* (indice di massa corporea) e *Life_expectancy* (aspettativa della vita in anni)

Indaghiamo quanto varia l'aspettativa di vita nelle macroregioni del mondo considerate. Nella *Fig.3* notiamo che le varie macroregioni presentano range di variazione e mediane diversi. Le macroaree con

l'aspettativa di vita più alta sono l'Europa e l'America settentrionale, con una mediana intorno agli 78 anni. Le macroaree con l'aspettativa di vita più bassa sono l'Africa con una mediana intorno ai 57 anni. Le altre macroregioni si collocano tra i 65 e i 75 anni.

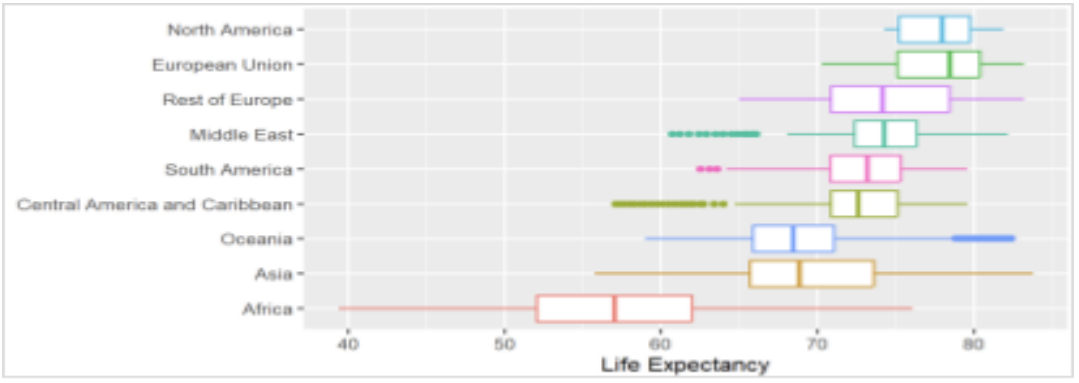

Fig. 3: Boxplot di aspettativa di vita per ogni macroregione.

Le macroregioni con la maggiore varietà sono l'Africa e Asia con range di 35 e 30 anni rispettivamente, questo significa che ci sono paesi molto diversi tra loro in termini di aspettativa di vita.

Nel grafico sottostante (*Fig.4*) vediamo come le densità della variabile *Life expectancy* siano distribuiti all'interno delle classi di aspettativa di vita, vediamo che la distribuzione del secondo gruppo è poco centrata e tendente al terzo gruppo, temiamo che in fase di classificazione molti missclassified saranno proprio elementi di questi due gruppi.

Andando ad analizzare il comportamento di ogni variabile condizionatamente alle classe di interesse, nella *Fig.5* vediamo che ci sono alcune variabili che si comportano diversamente in tutte le classi di appartenenza come ad esempio *Schooling* e *Thinness_fife_nine_years* e quindi discriminano bene le classi. Variabili come *Dipthteria*, *GDP_per_capita* e *Alcohol consumption* sembrano avere differenze non tra tutte le classi, ma solo tra alcune di esse.

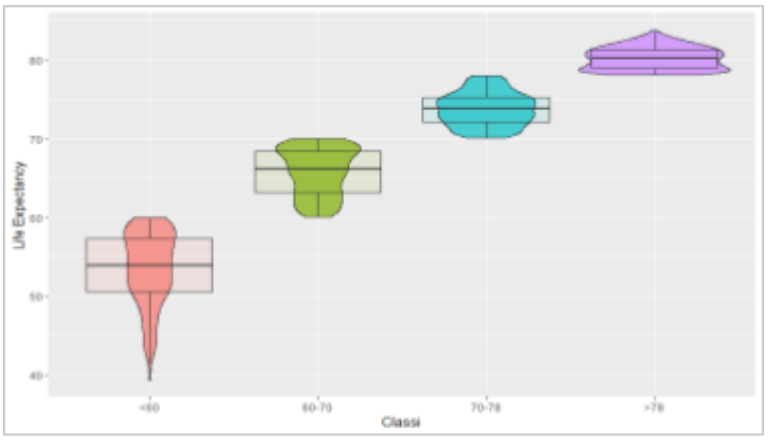

Fig. 4: Grafico a violino con boxplot sovrapposto

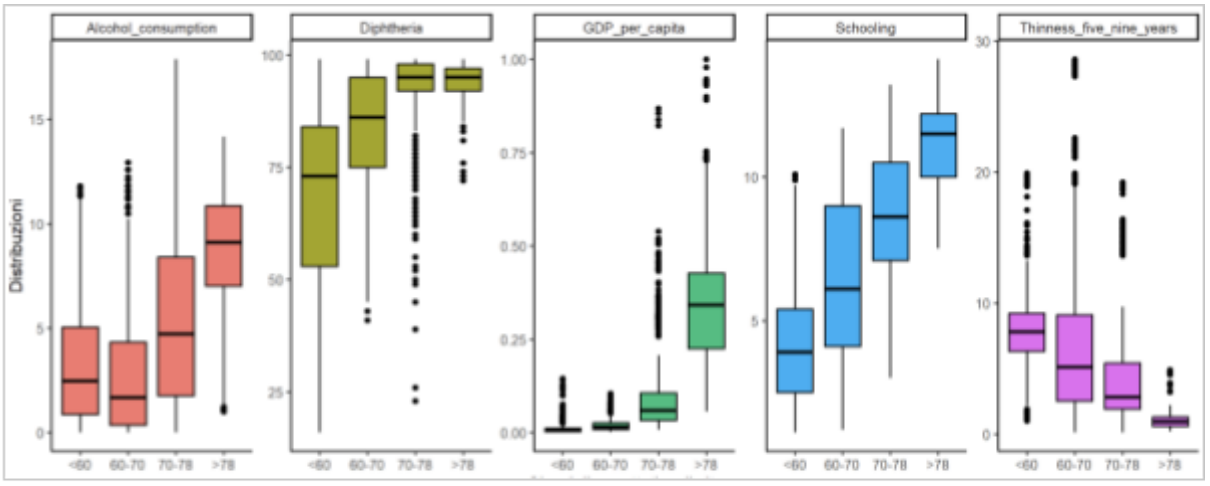

Fig. 5: Boxplot relativi alle cinque variabili *(Alcohol consumption, Diphtheria, GDP per capita, Schooling e Thinnes 5-9)* ripartiti per ogni classe di aspettativa di vita.

4. MODELLI DI CLASSIFICAZIONE SUPERVISIONATA

Per proseguire con i due metodi di classificazione scelti, MDA e EDDA abbiamo suddiviso il dataset in *training* e *test* set:

- in training set abbiamo deciso di conservare le statistiche che sono state raccolte tra 2000 e 2014;
- in test i dati sono invece per il 2015. Vogliamo testare la qualità della previsione dei modelli utilizzando come nuove osservazioni l'anno più recente.

Vogliamo valutare quale dei due metodi di classificazione MDA e EDDA produce un modello con l'accuracy migliore. Per ognuno dei due metodi sono stati implementati 14 modelli basati sulla distribuzione gaussiana. La classificazione viene suddivisa in fase di *learning* dove implementiamo, valutiamo e selezioniamo i modelli e fase di *test*

MODELLO	CV	BIC	ACCURACY TEST
EVE	**0,24**	113837,80	0,73
EEE	0,25	126344,00	**0,75**
VVV	0,24	**111442,00**	0,72

Tab. 1: Tre migliori modelli EDDA con valori di CV BIC e l'Accuracy della previsione.

Valori veri \ Classificazione	<60	70-70	70-78	>78	
<60	12	7	0	0	19
70-70	8	29	8	0	45
70-78	0	8	71	1	80
>78	0	0	12	23	35
	20	44	91	24	179

Tab. 2: EDDA confusion matrix

Valori veri \ Classificazione	<60	70-70	70-78	>78	
<60	18	1	0	0	19
70-70	17	22	6	0	45
70-78	0	13	67	0	80
>78	0	0	9	26	35
	35	36	82	26	179

Tab. 3: MDA confusion matrix

dove valutiamo la capacità di assegnazione delle classi dei modelli scelti.

Nella fase di learning per EDDA la stima del modello è stata ripetuta più volte in modo da tener conto della aleatorietà della stima, i modelli migliori dal punto di vista di CV e BIC sono riassunti nella *Tab. 1* Il migliore risulta essere EVE (gruppi con stesso volume e orientamento ma forme diverse), ma dopo aver confrontato i valori di accuracy nella fase di previsione, il modello EEE (stesso volume, forma e orientamento) risulta fornire le previsioni con accuracy pari a *75,41%*, il che ci porta a sceglierlo come modello migliore dato anche il vantaggio della complessità minore rispetto EVE.

Il miglior modello nella stima con metodo MDA risulta essere composto da quattro misture stimata con i seguenti modelli: EEV, VVE, VEV ,VVV rispettivamente per ogni fascia di aspettativa di vita. Ogni modello è composto a sua volta da 5 componenti.

Nella fase di previsione il modello MDA scelto risulta prevedere le "nuove" osservazioni (test set) con un'accuracy pari a *74,3%.*

Nelle matrici di confusione (*Tab. 2, Tab. 3*) possiamo vedere più in dettaglio come i nostri modelli di classificazione assegnano le osservazioni ai vari gruppi.

Di seguito vediamo come i paesi vengono classificati e che differenze ci sono tra le classi dei nostri modelli e quelle reali.

Come possiamo osservare da questi grafici in entrambi le classificazioni non sono presenti differenze significanti rispetto alla cartina originale, tuttavia visivamente notiamo un adattamento leggermente migliore del modello EDDA rispetto al modello MDA.

Fig 7: Mappa del mondo previsione della classe per l'aspettativa di vita modello EDDA.

Fig. 8: Mappa del mondo previsione della classe per l'aspettativa di vita modello MDA.

5. CLUSTERING

Nella parte di clustering abbiamo dapprima cercato quanti gruppi fornissero la verosimiglianza più alta, verificando se l'algoritmo cogliesse gli stessi gruppi che noi abbiamo ipotizzato o in caso contrario capire quale associazione creasse in autonomia. Abbiamo notato in un primo ciclo che il numero di gruppi del modello migliore era 9, abbiamo per prima cosa verificato se i gruppi che trovava l'algoritmo fossero per caso le macroregioni (anch'esse 9), ma non era così. Infatti ha prodotto un error rate molto alto ed abbiamo notato che se aumentando i gruppi oltre la soglia di default, l'algoritmo sceglieva il numero massimo di gruppi disponibili, andando in overfitting, segno che non riusciva a trovare dei gruppi all'interno del nostro dataset.

Forzando l'algoritmo a scegliere un numero di gruppi pari alle classi della nostra variabile d'interesse il modello migliore risulta essere un modello VVV. Tuttavia questo modello produce un error rate del 35%, non abbastanza soddisfacente per la nostra analisi. Di seguito il grafico delle osservazioni clusterizzate, con i rispettivi errori (puntini neri).

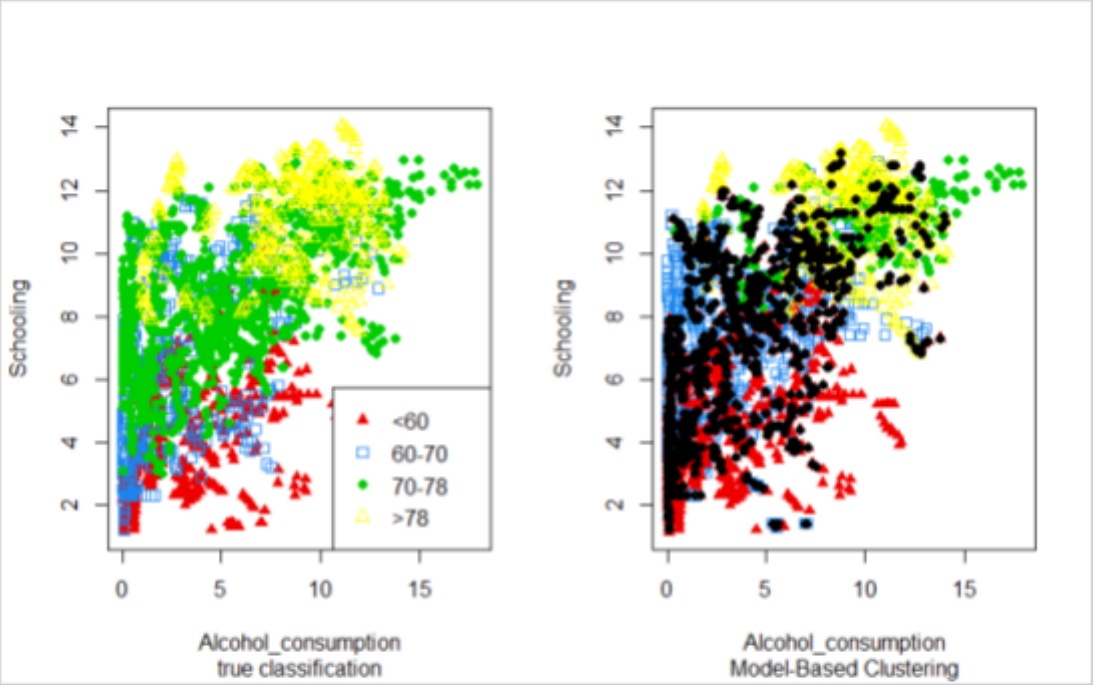

Fig. 9: Scatterplot di variabili Alcohol e Scholling. Nel grafico a sinistra sono presenti le vere classi.Nel grafico a destra è rappresentata la nostra clusterizzazione, con punti neri sono stati segnati i paesi assegnati al gruppo "sbagliato".

6. CONCLUSIONI

Dalle nostre analisi abbiamo notato che alcune variabili non scontate sono positivamente correlate con l'aspettativa di vita, queste variabili sono *Schooling, GDP_per_capita, Diptheria e BMI*.

La classificazione ci ha portato dei buoni risultati creando una mappa coerente con la reale aspettativa di vita inerente all'anno 2015. Alcuni degli errori derivanti dai nostri modelli possono sicuramente essere ricondotti alla suddivisione fatta a priori della variabile *Life expectancy.* Tuttavia da questi modelli non abbiamo riscontrato nessun tipo di errore "grossolano", le osservazione classificate non correttamente sono state assegnate a fasce di aspettative di vita "vicine" alla reale classe.

Valutando le confusion matrix di ognuno dei due modelli, *Tab. 2* e *Tab 3* si nota come modello scelto EDDA ha delle prestazioni migliori a classificare i paesi che hanno aspettativa di vita tra 70 e 78 anni.

Invece la classificazione con MDA è molto performante per i paesi con aspettativa di vita maggiore di 70 anni, mentre non riesce a cogliere bene la classe di paesi con l'aspettativa di vita meno di 60 anni. Il modello migliore risulta per noi essere quello EDDA, sia come prestazioni, sia come semplicità.

Discorso diverso è stata la clusterizzazione, la quale non riesce a cogliere dei gruppi all'interno del dataset e a produrre buoni risultati, imponendo il numero di cluster uguale al numero di classi della variabile di interesse il risultato migliore che otteniamo è una precisione del 65%, inferiore a quella della classificazione.

APPENDICE

Link dataset: https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated)