Country Statistics

UNData

The Outliers

Giulio Pedicone

Vincenzo Villanova

Francesco Virzì P. A.

Università di Catania

Cosa determina lo sviluppo di una nazione?

I dati utilizzati in questo progetto sono disponibili sulla piattaforma

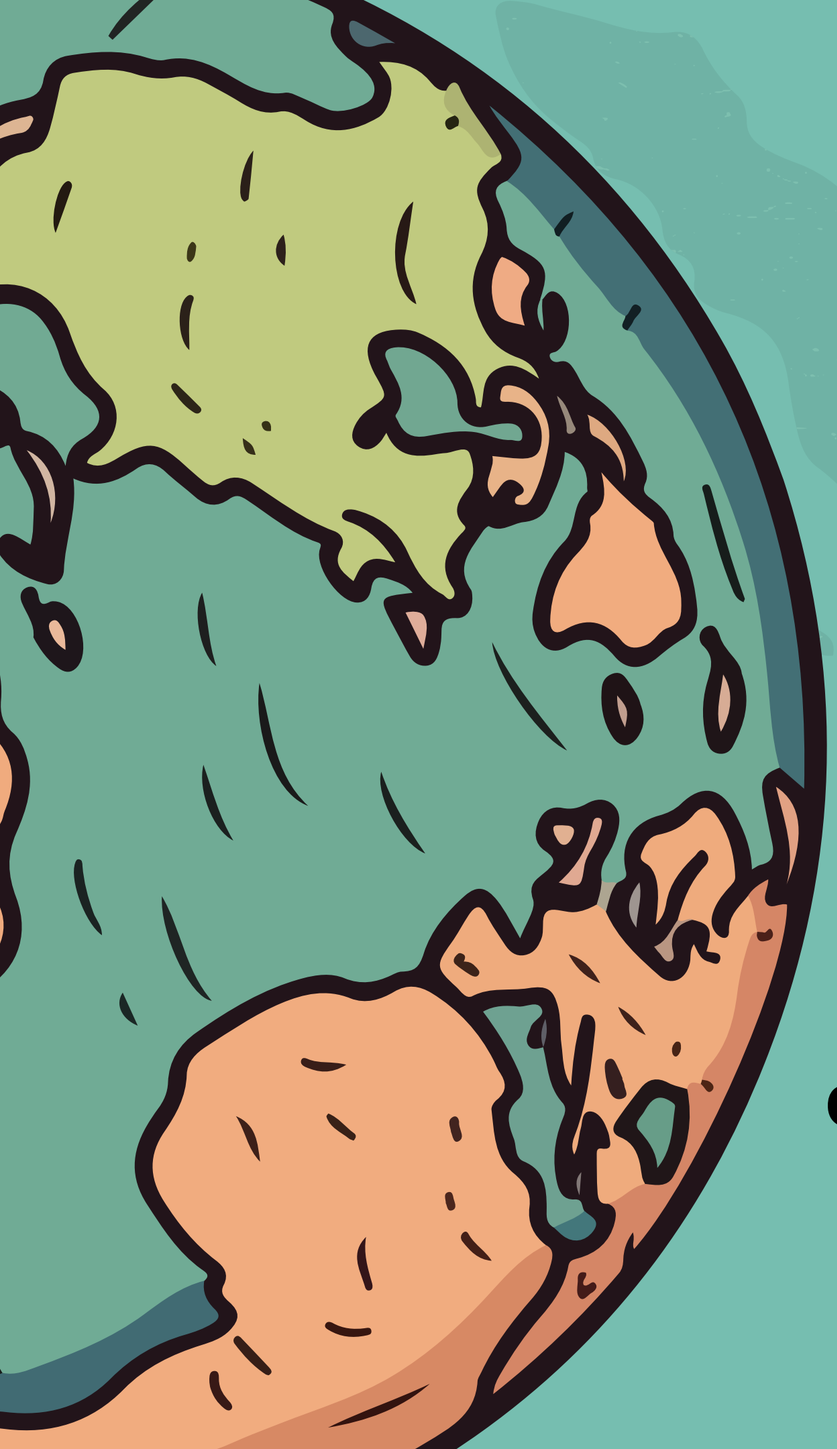
kaggle

La Sfida del Data Cleaning

2
files CSV

239
osservazioni

50
features

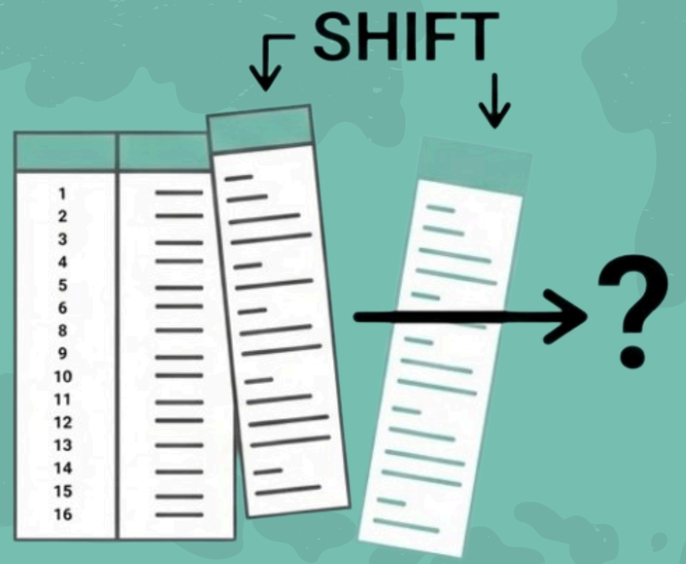


Inconsistenze nel dataset

- Osservazioni **duplicate**
- Feature composte da **2 variabili** (es: ratio f/m per 100 pop.)
- **Valori mancanti** da gestire correttamente (-99 o ...)
- **Colonna duplicata** (Mobile Cellular Subscriptions)
- **Shift** dei dati nelle relative colonne

230
osservazioni

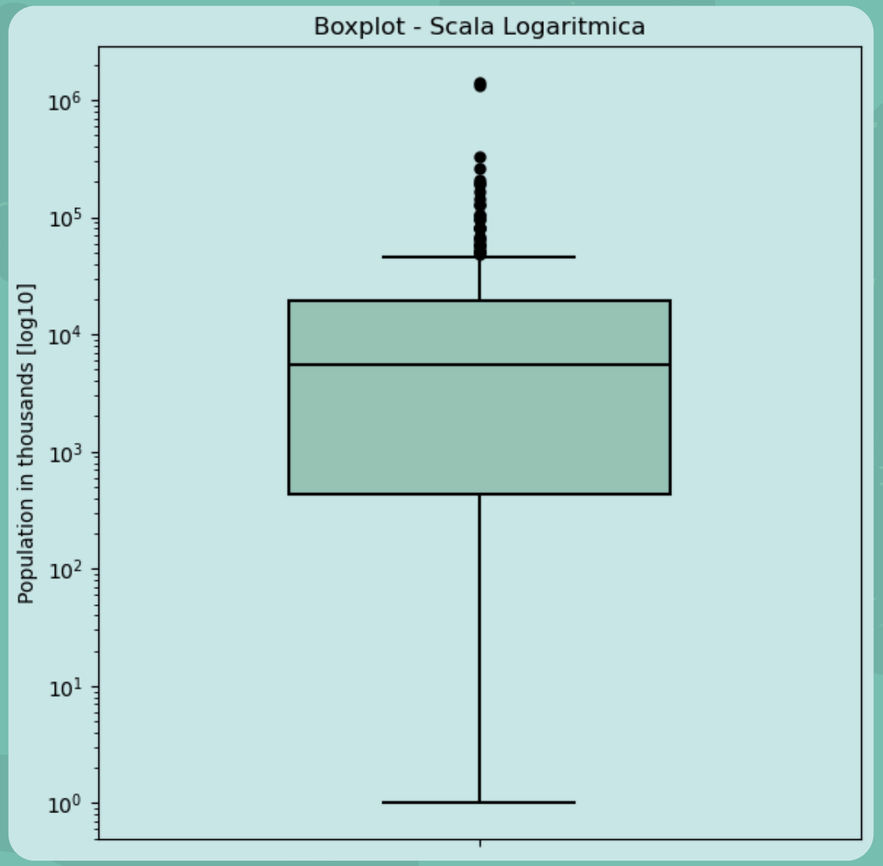
59
features



Dati anomali?

Non sono davvero degli outliers!

Riflettono la reale eterogeneità dei dati nei diversi paesi del Mondo.



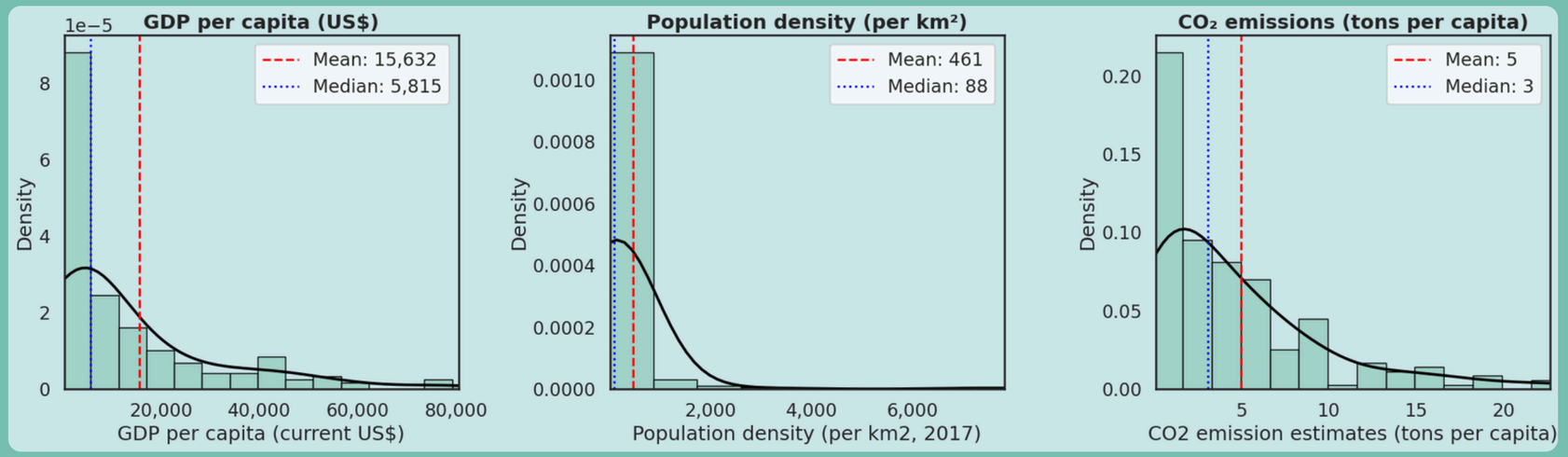
Analisi Univariata

- **GDP** per capita
- **Densità** della **Popolazione**
- **Emissioni** di **CO2**

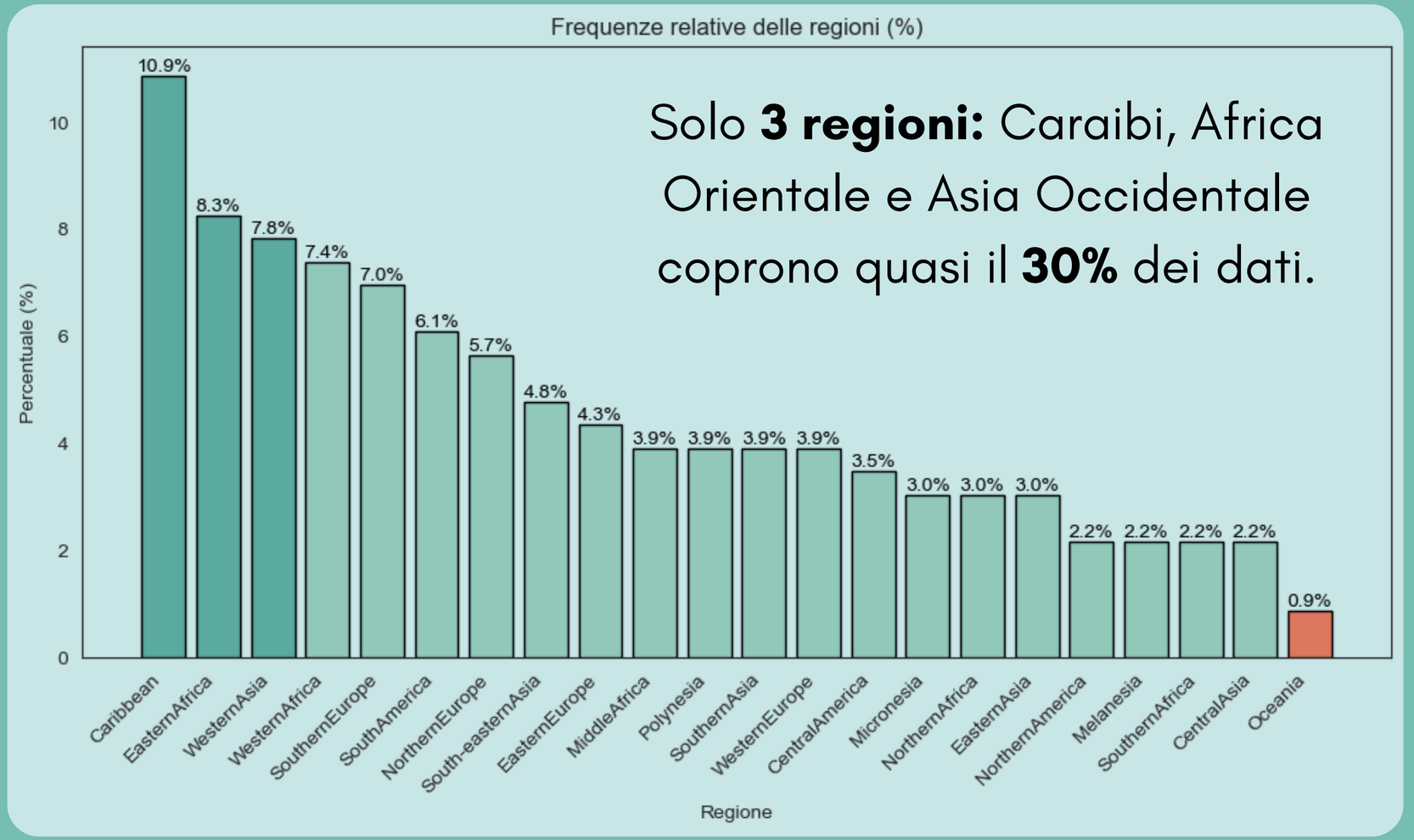
Marcata Asimmetria Positiva

↓

Forte Disuguaglianza/Disomogenità



Come sono distribuiti i dati?



Solo **3 regioni:** Caraibi, Africa Orientale e Asia Occidentale coprono quasi il **30%** dei dati.

Distribuzioni così sbilanciate influenzano i risultati dei modelli

Il Nucleo dello Sviluppo (Analisi Multivariata)

Forte correlazione tra:

- **GDP** per Capita
- **Emissioni** di **CO2**
- Utilizzo di **Internet**

Correlazioni Spearman — Evidenza ρ ≥ 0.8

	CO2 emission estimates (tons per capita)	GDP per capita (current US$)	Urban population (% of total population)	Population density (per km2, 2017)	Individuals using the Internet (per 100 inhabitants)
CO2 emission estimates (tons per capita)	1.00	**0.87**	0.66	0.06	**0.82**
GDP per capita (current US$)	**0.87**	1.00	0.68	0.16	**0.91**
Urban population (% of total population)	0.66	0.68	1.00	0.02	0.65
Population density (per km2, 2017)	0.06	0.16	0.02	1.00	0.22
Individuals using the Internet (per 100 inhabitants)	**0.82**	**0.91**	0.65	0.22	1.00

Salute e Benessere

Abbiamo predetto tramite **regressione lineare** la **mortalità infantile** ogni **1000 nascite**

91.8%

varianza spiegata

-2.34
Aspettativa di vita

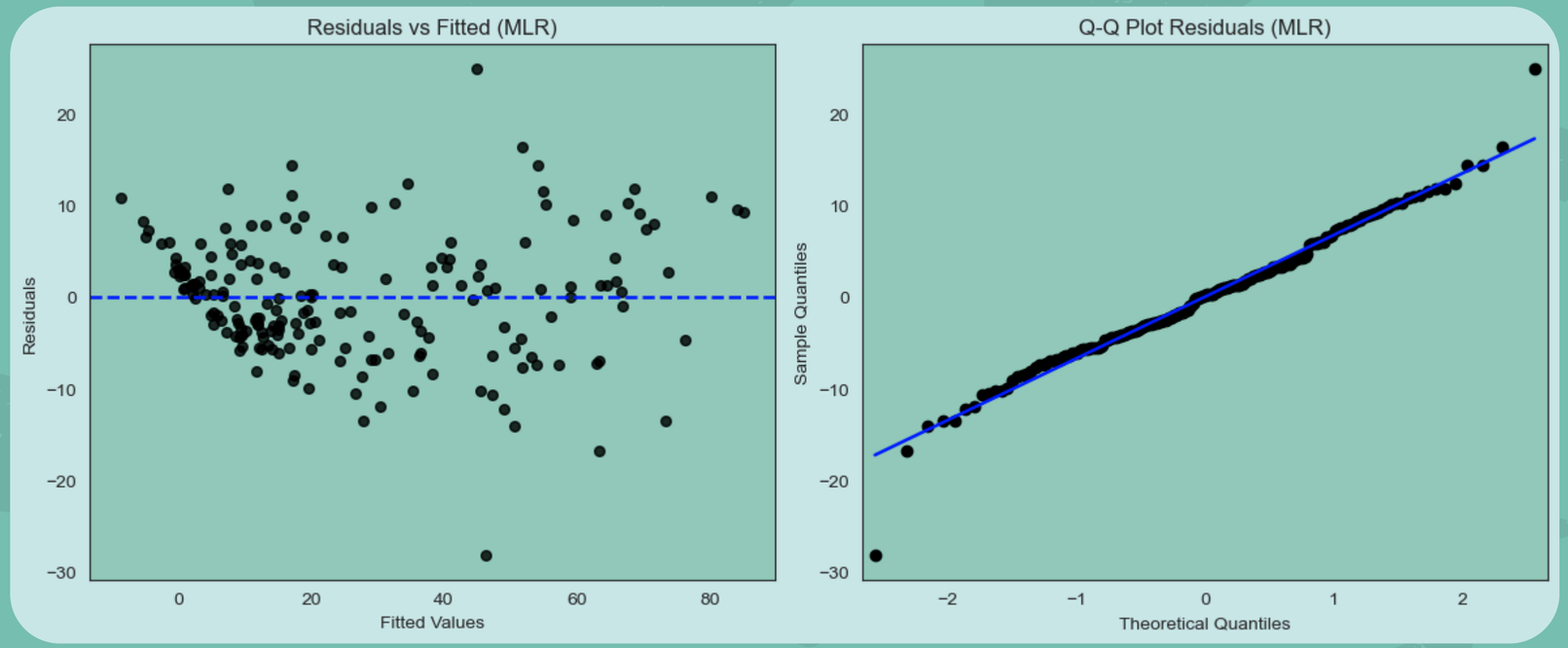
-0.23
Emissioni di CO2

-0.16
Utilizzo di Internet

Salute e Benessere

Distribuzione a "**imbuto**" dei punti rivela **eteroschedasticità**,
La varianza degli errori non è costante ma **aumenta** al crescere dei valori previsti.

Il **distacco** dei punti dalla retta alle estremità indica che i residui **non sono distribuiti normalmente**, segnalando la presenza di **outlier.**



Qual è la vera ricchezza?

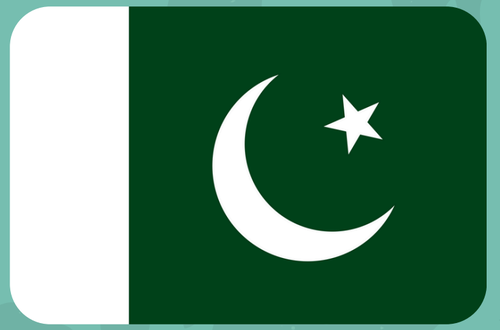
Il nostro **classificatore binario** spiega correttamente il **74%** della probabilità che un Paese sia ricco.

+32%
Aspettativa di vita

+12%
Utilizzo di Internet

-23%
Mortalità infantile

Italia
99.9%
di probabilità di essere un **paese ricco**



Pakistan
0%
di probabilità di essere un **paese ricco**

Machine Learning: Considerazioni

Dati limitati

il numero di osservazioni non era sufficiente per predizioni affidabili.

Scelta metodologica

è stata adottata la Cross-Validation con K-Fold.

Risultato

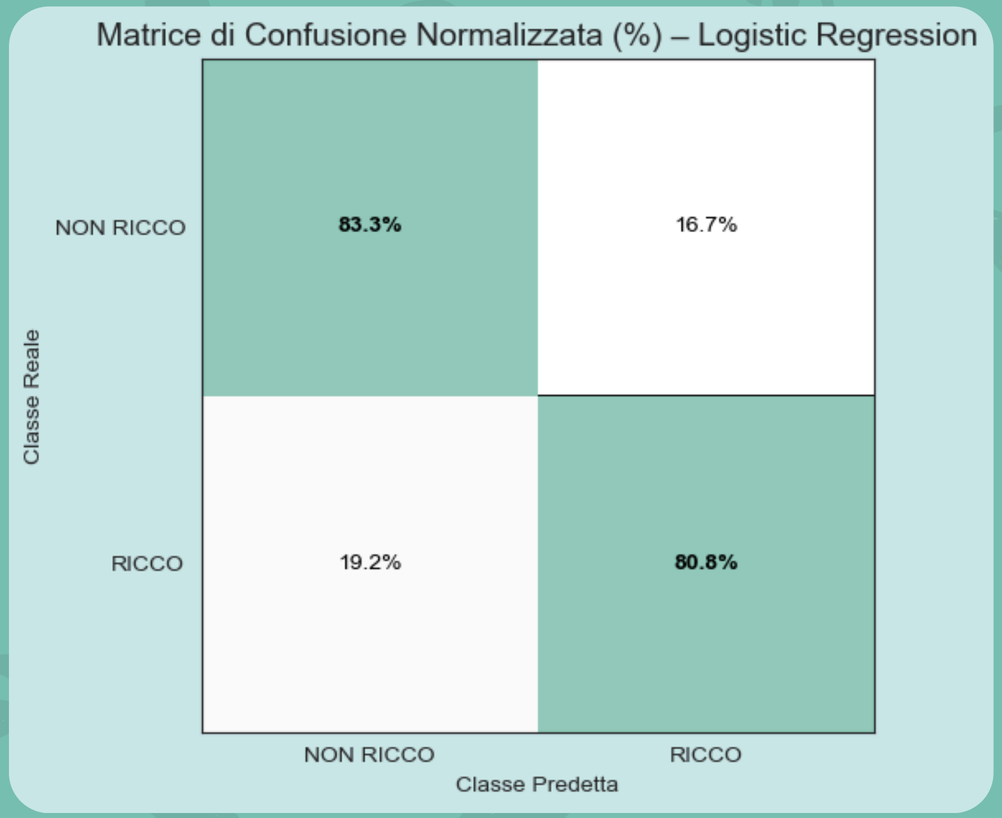
valutazioni più robuste e stabili delle performance del modello.

Classificazione: Approccio ML

Abbiamo utilizzato una **regressione logistica** per classificare i Paesi in ricchi e non ricchi sulla base di indicatori **socio-economici**, definendo il target tramite la mediana del **PIL pro capite**.

82%
accuracy

80%
precision

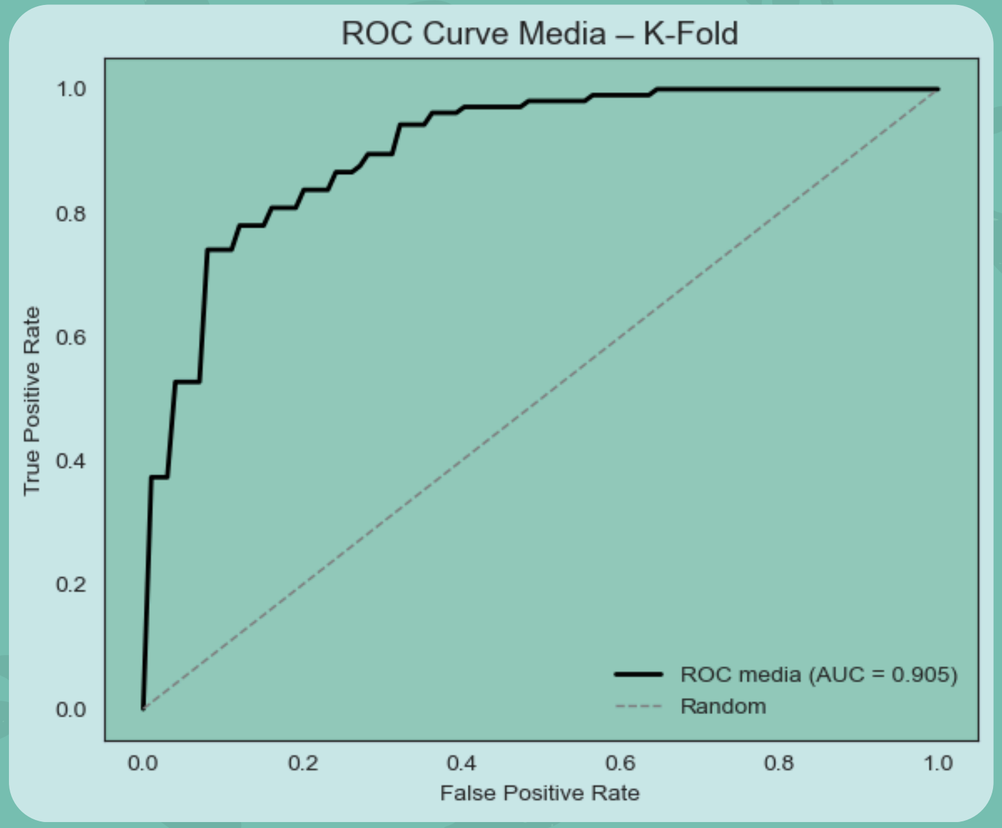


Classificazione: Approccio ML

Abbiamo utilizzato una **regressione logistica** per classificare i Paesi in ricchi e non ricchi sulla base di indicatori **socio-economici**, definendo il target tramite la mediana del **PIL pro capite**.

82%
accuracy

80%
precision

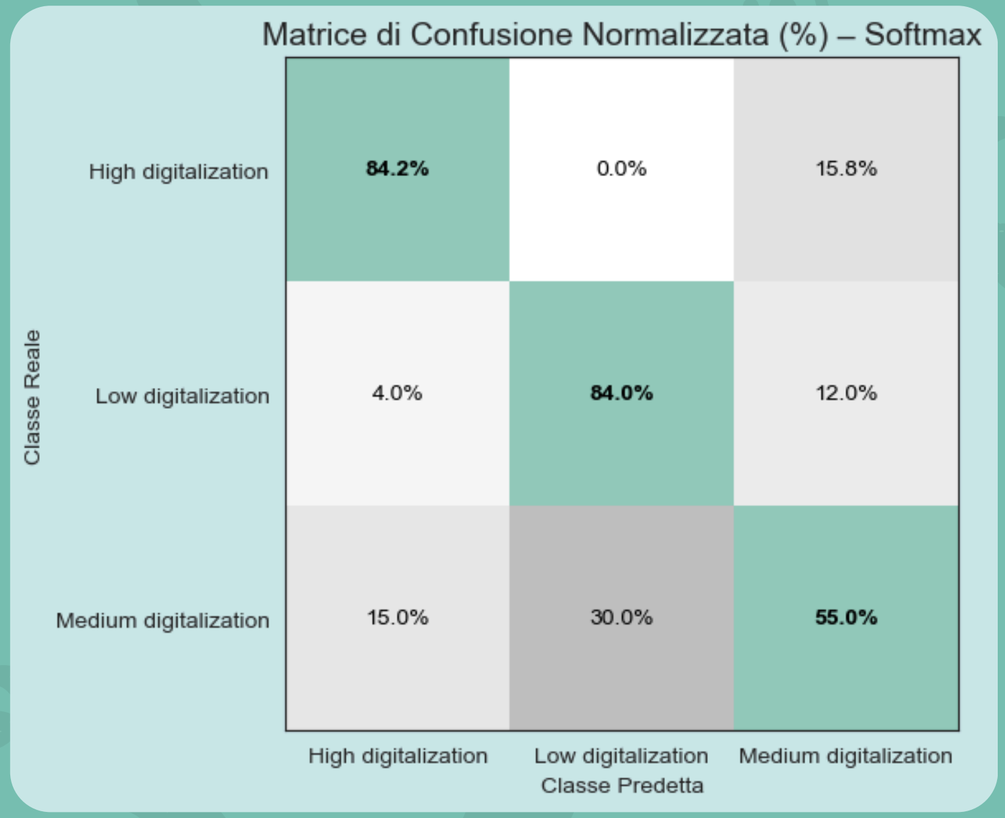


Classificazione: SoftMax

Tramite una **regressione logistica multinomiale** (Softmax) abbiamo classificato i Paesi in **tre livelli di digitalizzazione**, definiti a partire dall'uso di Internet e predetti tramite indicatori socio-economici.

75%
accuracy

74%
precision

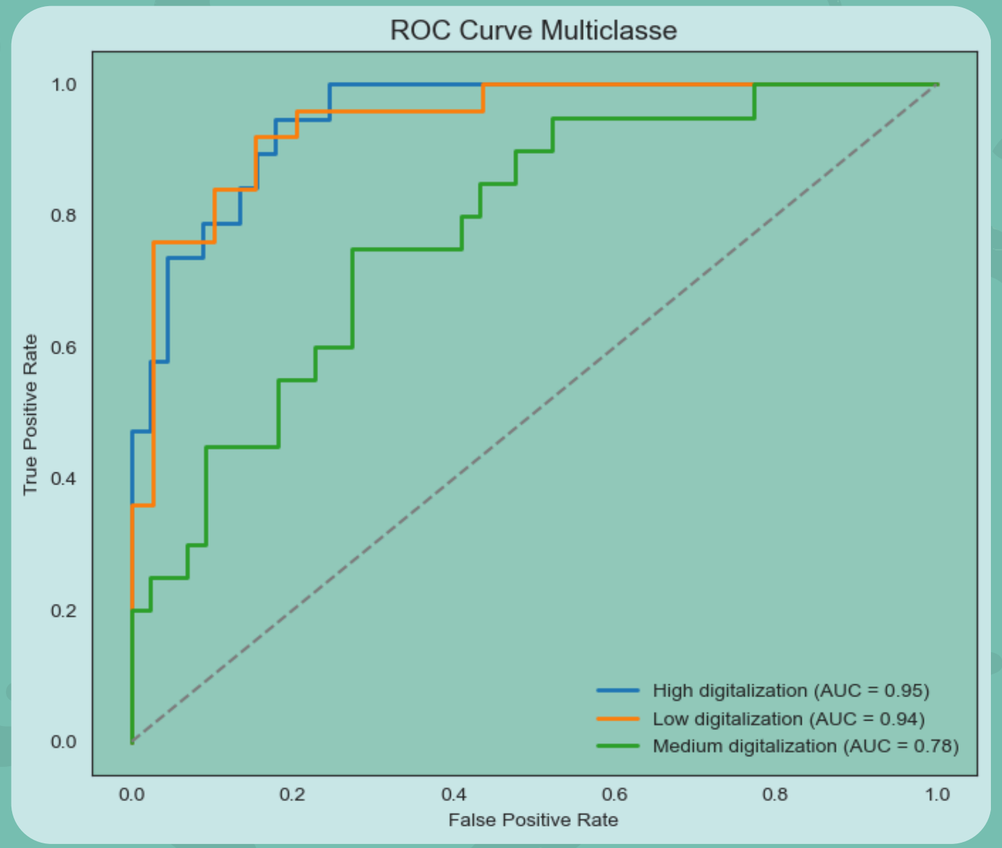


Classificazione: SoftMax

Tramite una **regressione logistica multinomiale** (Softmax) abbiamo classificato i Paesi in **tre livelli di digitalizzazione**, definiti a partire dall'uso di Internet e predetti tramite indicatori socio-economici.

75%
accuracy

74%
precision



Confronto con One VS Rest

Il confronto con la versione **One-vs-Rest** ha evidenziato un **degrado complessivo** delle prestazioni.

75% accuracy → **67%** accuracy

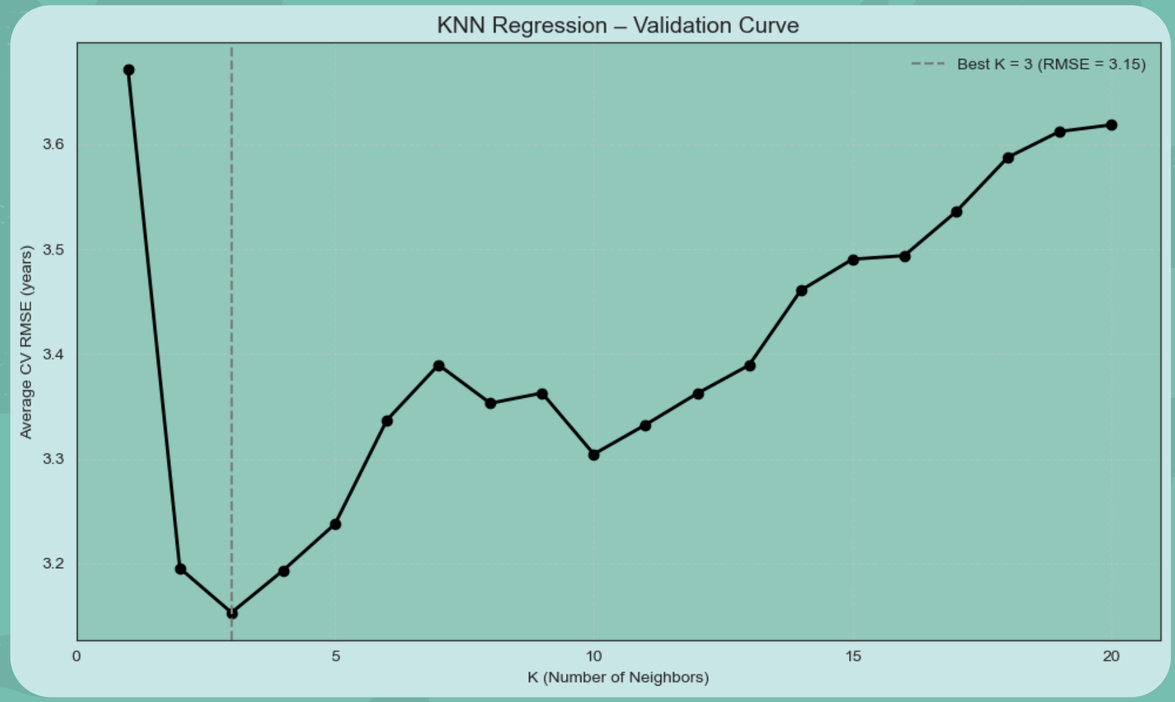
74% precision → **65%** precision

Regressione: K-Nearest Neighbor

KNN stima l'**aspettativa di vita media** di un Paese da indicatori socio-economici e sanitari, catturando relazioni non lineari. Utile per confronti internazionali e supporto a politiche sanitarie.

K = 3

Miglior numero di classi



Cosa abbiamo scoperto?

Le **stime dell'aspettativa** di **vita media** sono accurate ed hanno **metriche ragionevoli.**

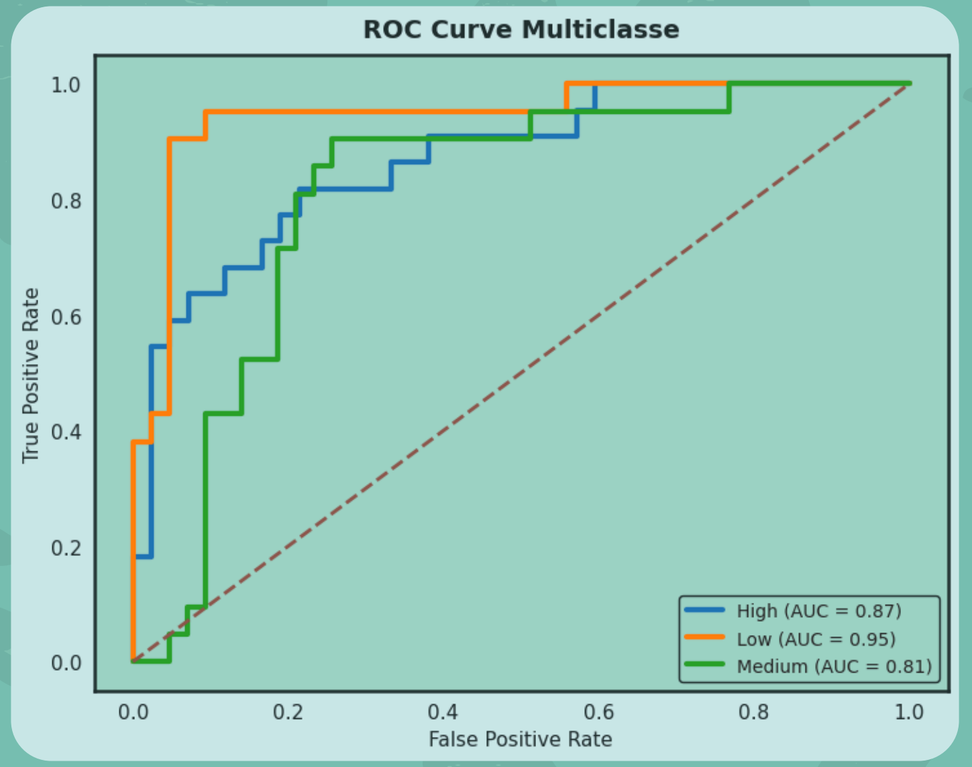
MAE: 2.89 anni **RMSE:** 4.47 anni

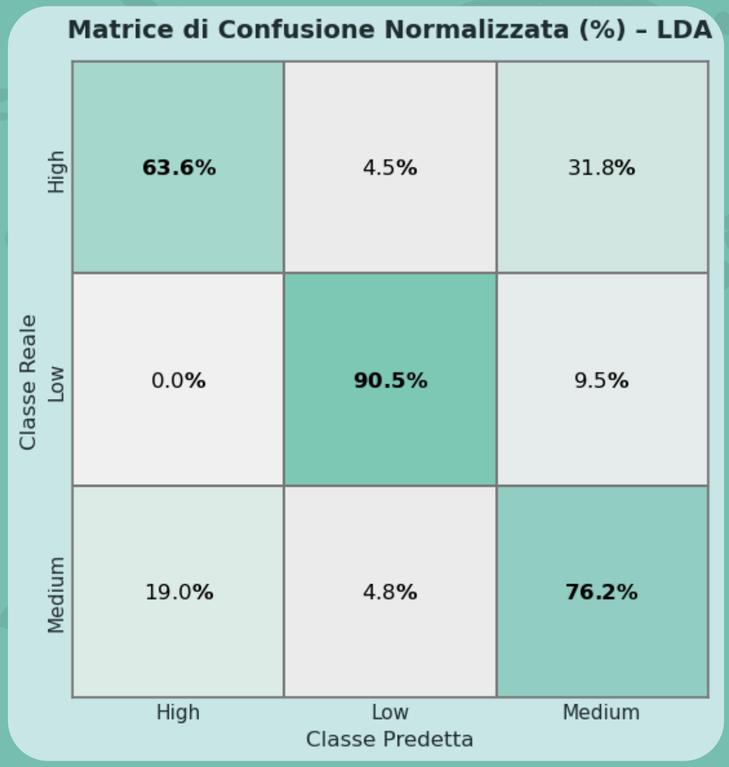
San Marino
(84 anni)

Repubblica Centrafricana
(49 anni)

Classificazione: LDA

La Linear Discriminant Analysis classifica i Paesi in base **all'aspettativa di vita** (**bassa**, **media** o **alta**) utilizzando indicatori socio-economici, creando confini di decisione lineari facilmente interpretabili.





Classificazione: LDA

La classe meglio riconosciuta è quella con **basse** aspettative di vita con **precision** e **recall** pari al:

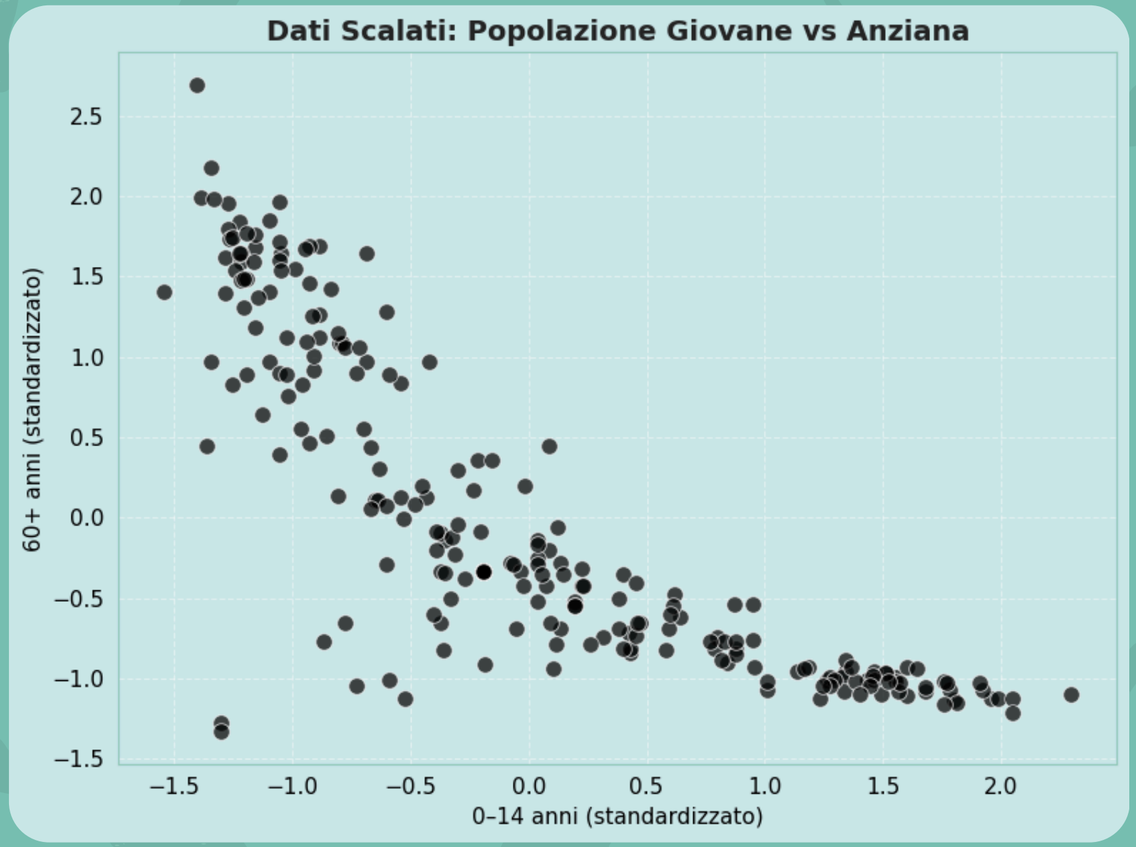
90%

La classe con aspettativa di vita **media** è più difficile da distringue con **recall** pari al:

76%

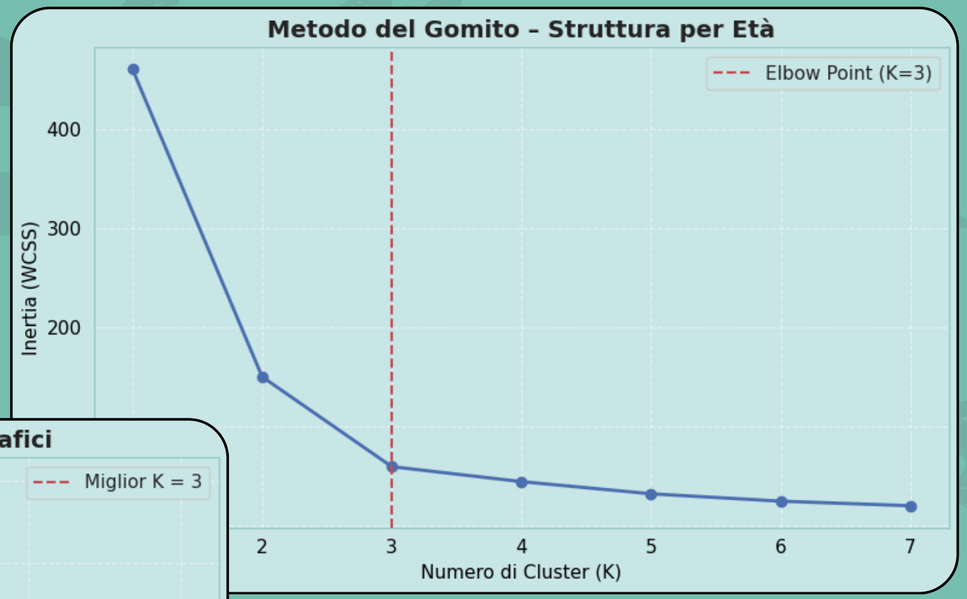
Gruppi Demografici: Clustering

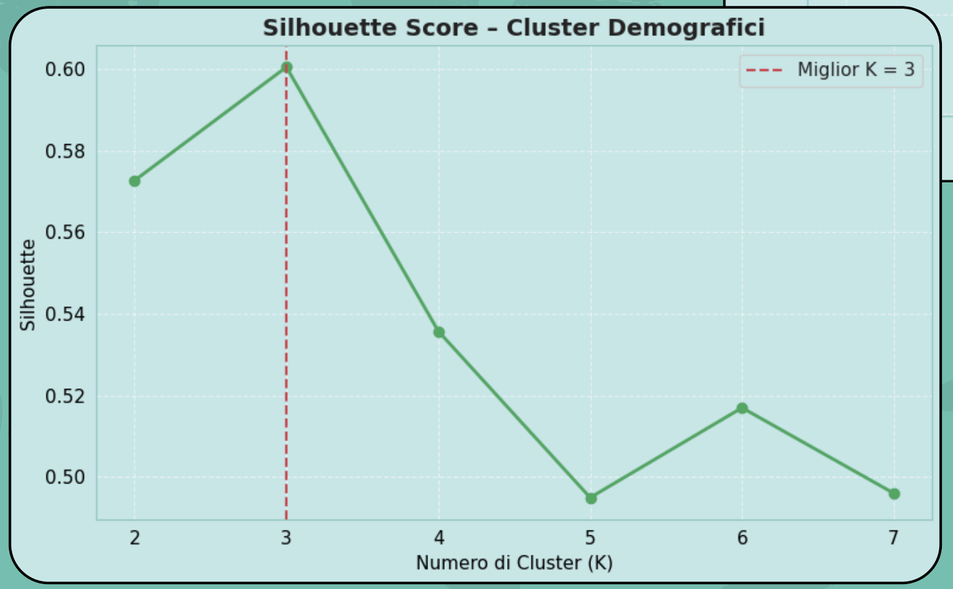
È stato applicato un **clustering** sui Paesi, per individuare fasi della transizione demografica. Per capire meglio le differenze nella struttura della popolazione e identificare pattern comuni usiamo solo indicatori demografici (quota 0-14 e 60+).



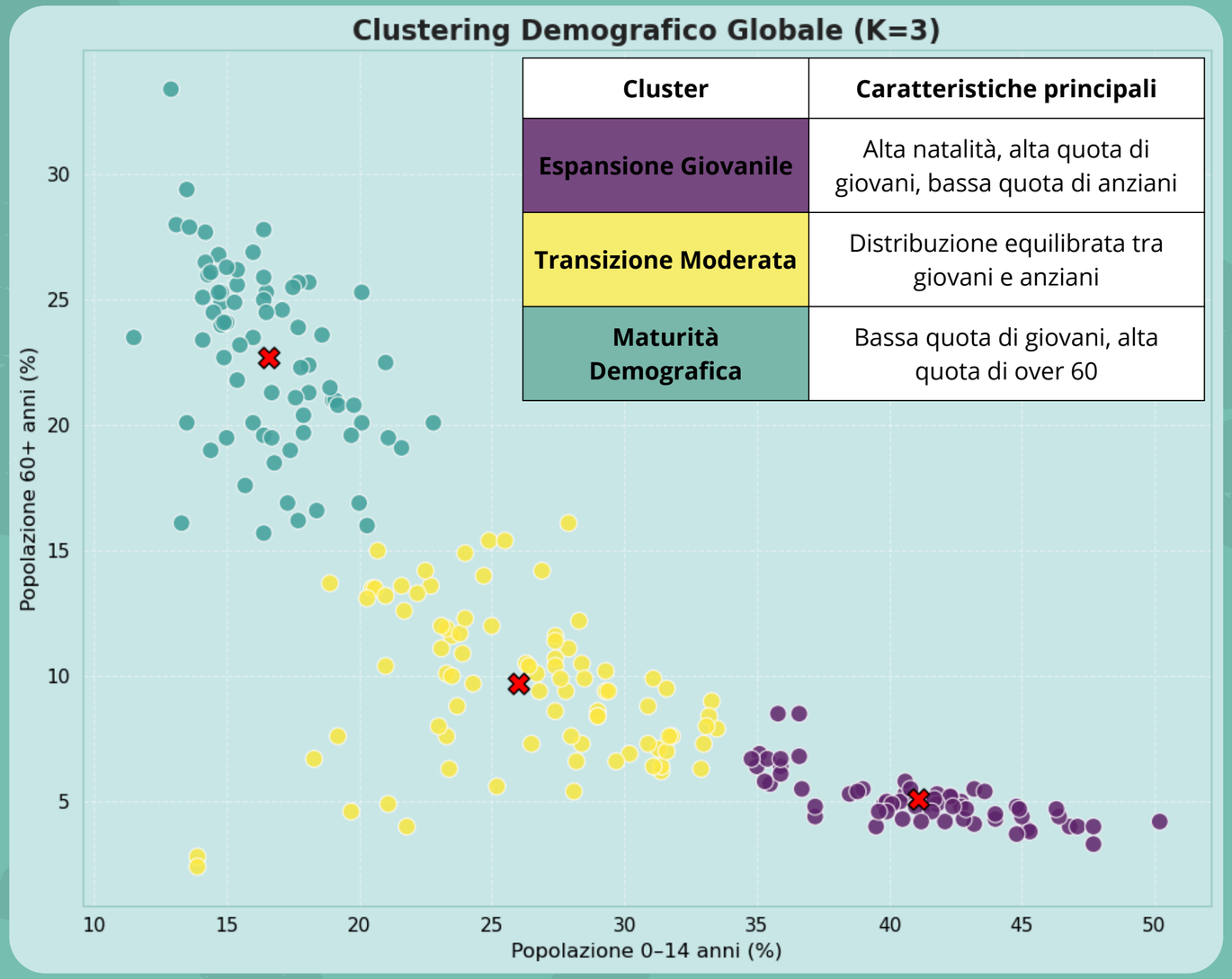
Clustering

Con K-Means sono stati individuati **3** cluster demografici





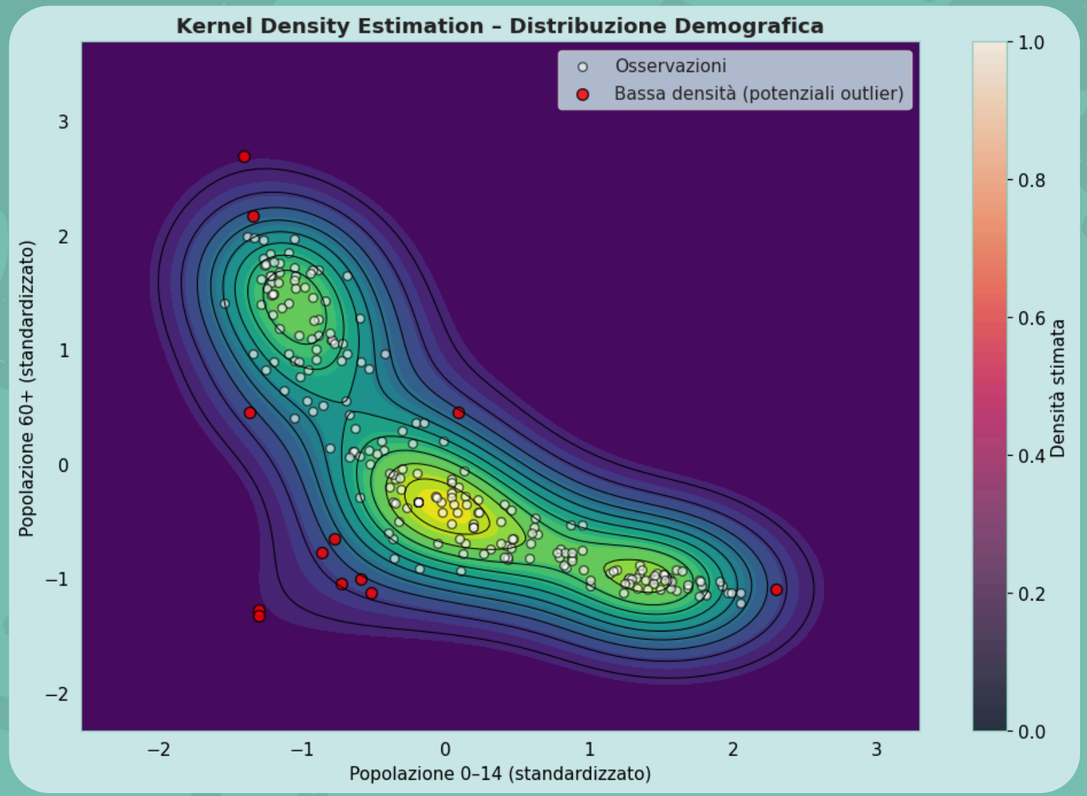
3 Cluster Interpretabili



Cluster	Caratteristiche principali
Espansione Giovanile	Alta natalità, alta quota di giovani, bassa quota di anziani
Transizione Moderata	Distribuzione equilibrata tra giovani e anziani
Maturità Demografica	Bassa quota di giovani, alta quota di over 60

Stima di Densità

La **KDE** evidenzia aree ad alta densità **coerenti con i tre cluster demografici**. Rispetto allo scatter plot, fornisce una visione più continua e informativa.



Oltre la Statistica: Prevedere lo Stato di Sviluppo

2015

62

Aspettativa di vita

2015

0.45%

di probabilità di essere un **paese ricco**

2020

64.4

Aspettativa di vita

2020

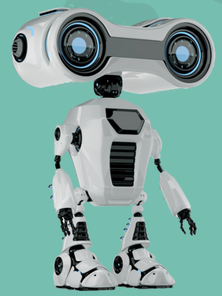
88.8%

di probabilità di essere un **paese ricco**

Conclusioni

Lo sviluppo di una nazione non è un numero, ma un **ecosistema interconnesso**.
La vera ricchezza è il risultato di **tre** pilastri:

Salute



Tecnologia

Sostenibilità

GRAZIE

The Outliers

Q&A SESSION