Università di Pisa

Corso di Laurea in

Artificial Intelligence and Data Engineering

Classificazione della politica di restrizione sugli eventi pubblici in base ai dati della pandemia da COVID-19

Candidato

Franco Terranova

ANNO ACCADEMICO 2021/2022

INDICE

INTRODUZIONE

Si ipotizza che l'analisi sia stata commissionata da un'azienda che si occupa dell'organizzazione di tour mondiali di eventi musicali e concerti. Lo scopo dell'analisi è la classificazione della politica di restrizione sugli eventi pubblici nei vari Stati del mondo, esaminando l'andamento della pandemia da COVID-19. Infatti, alla luce dell'elevata volatilità dei nuovi casi, gli Stati potrebbero variare le proprie restrizioni sugli eventi pubblici e ciò comporterebbe una forte perdita economica da parte dell'azienda. Se una delle tappe del tour variasse le proprie restrizioni, sarebbe infatti necessaria una ri-organizzazione dell'evento. Il classificatore in questione dovrà raggiungere la migliore accuratezza possibile preferendo specificità a sensibilità, così da minimizzare il più possibile le perdite economiche.

1. DATASET

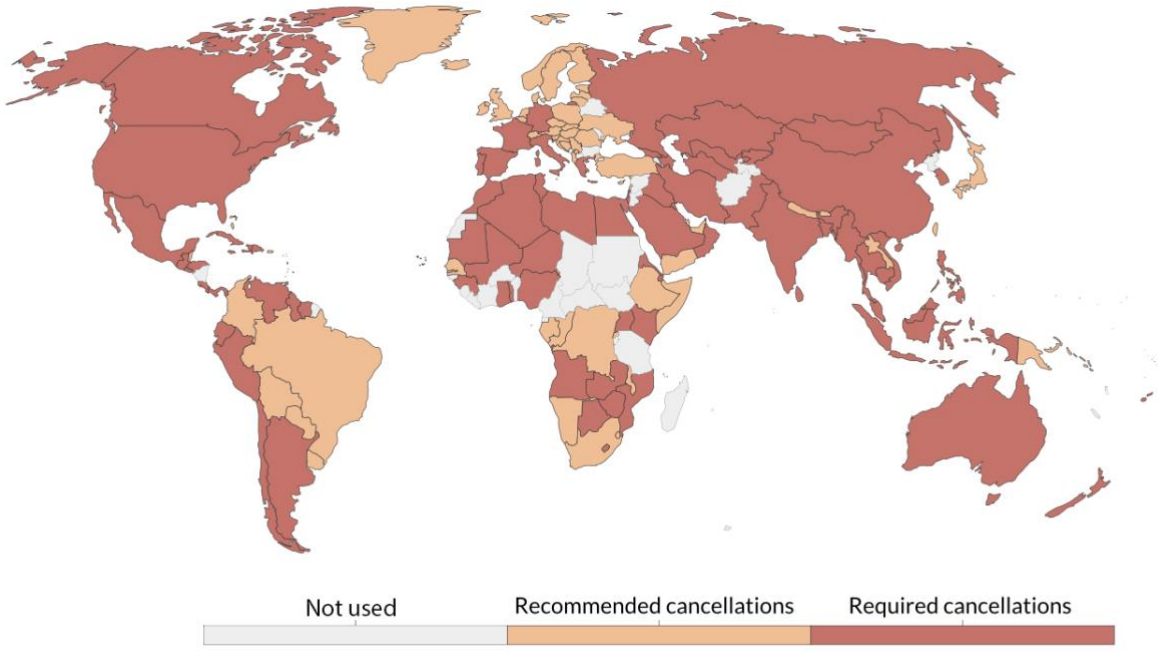
L'analisi è stata svolta sul dataset COVID-19 di Our World in Data, scaricabile al seguente link https://covid.ourworldindata.org/data/owid-covid-data.csv, e sul dataset relativo alla cancellazione di eventi pubblici, scaricabile dal seguente explorer https://ourworldindata.org/covid-cancel-public-events. I due dataset sono inoltre esaminabili attraverso il browser con il seguente Data Explorer https://ourworldindata.org/explorers/coronavirus-data-explorer.

1.1. CONTENUTO DELLA TABELLA

A partire dal dataset indicato è stata effettuata una prima operazione di pre-processing, come descritto nell'Appendice.
Successivamente, l'analisi è stata svolta sui seguenti fattori:

- stringency_index ($\in [0, 100]$): misura che riassume la risposta dei governi, ottenuta considerando diversi indicatori, tra cui chiusure di scuole, chiusure di posti di lavoro e divieti di viaggio, ridimensionati a un valore da 0 a 100 (100 = massima severità). Se le politiche variano a livello regionale, l'indice mostra la risposta della regione più rigorosa.
- total_cases ($\in \mathbf{Z}^+$): numero totale di casi di COVID-19 riscontrati finora.
- total_deaths ($\in \mathbf{Z}^+$): numero totale di morti da COVID-19 riscontrati finora.
- people_fully_vaccinated ($\in \mathbf{Z}^+$): numero totale di persone che hanno completato il ciclo vaccinale (due dosi).
- people_vaccinated ($\in \mathbf{Z}^+$): numero totale di persone che hanno ricevuto almeno una dose di vaccino.

- population ($\in \mathbf{Z}^+$): numero totale di persone che vivono nello stato.
- reproduction_rate ($\in \mathbf{Z}^+$): quante persone vengono contagiate in media da una persona infetta da COVID-19.
- new_cases ($\in \mathbf{Z}^+$): numero di nuovi casi giornalieri.
- new_deaths ($\in \mathbf{Z}^+$): numero di nuove morti giornaliere.
- public_events ($\in [0, 1]$): variabile categorica che vale 1 se non è obbligatoria la cancellazione di eventi pubblici, 0 se la cancellazione risulta essere obbligatoria. Potrebbero esserci differenze regionali nelle politiche sulla cancellazione degli eventi. In tal caso, uno Stato viene indicato come "obbligatorio" se almeno una regione richiede le cancellazioni.



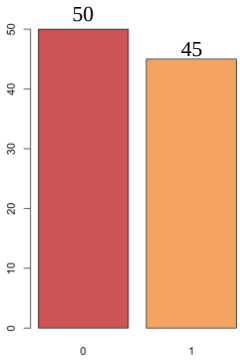
Sono stati utilizzati i dati relativi all'inizio di settembre, per i quali sono presenti un numero minore di osservazioni con valori mancanti.

Non sono stati considerati gli Stati per i quali sono presenti valori mancanti per i fattori utilizzati.

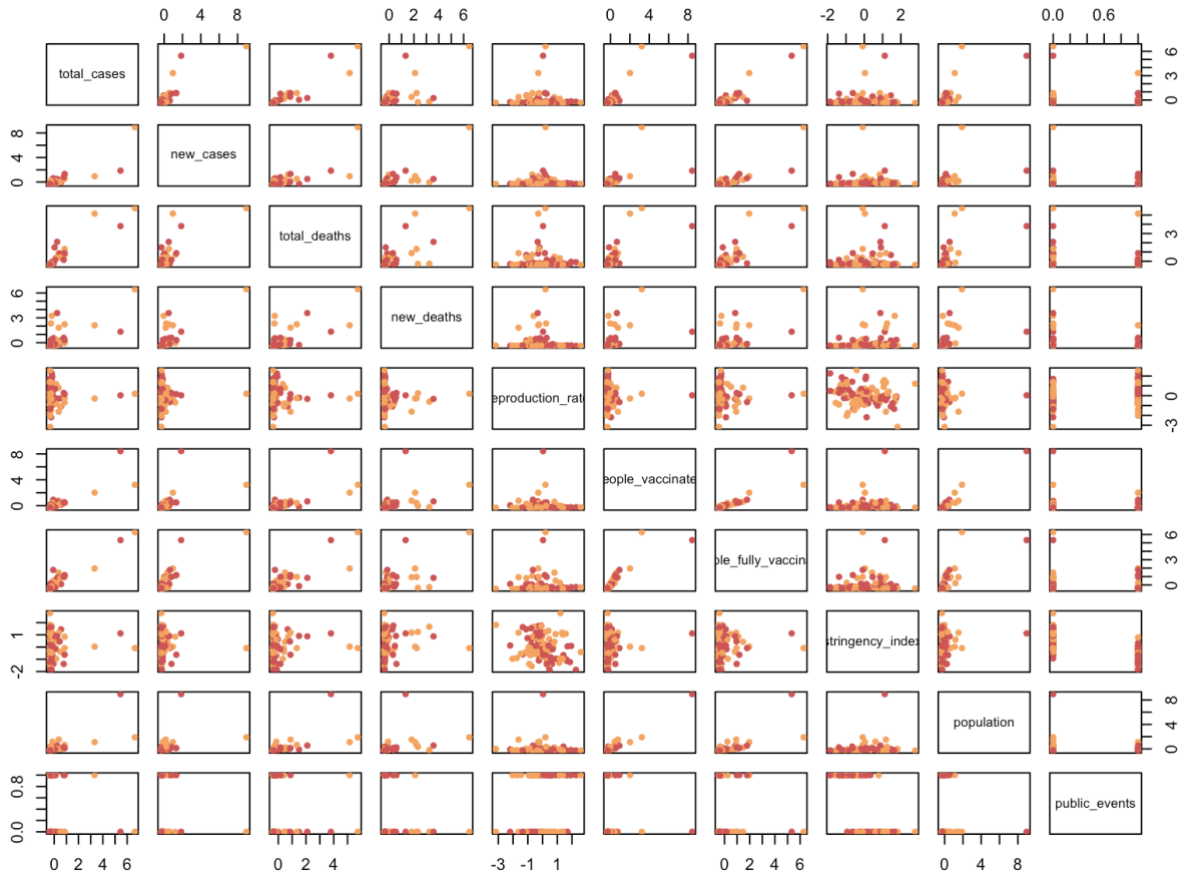
Inoltre, non si sono presi in considerazione gli Stati relativi alla classe "No measures" vista la presenza di un solo Stato, tra quelli con valori non mancanti per i fattori precedentemente scelti, appartenente a questa classe.

La tabella risultante contiene 10 colonne per un totale di 95 osservazioni.

Il dataset, come mostrato nel barplot accanto, risulta essere sufficientemente bilanciato.



Vista la differenza di scala tra i vari fattori nella tabella, si è proceduto con la standardizzazione della stessa. Di seguito troviamo lo scatter plot della tabella, in seguito alla standardizzazione.



Dallo scatterplot è possibile notare che non è presente una distinzione netta tra le due classi per praticamente tutti i fattori.

2. CONFRONTO TRA I METODI DI CLASSIFICAZIONE

Confrontiamo adesso i vari metodi di classificazione utilizzando come metriche l'accuratezza, la matrice di confusione, la curva ROC, l'area sotto la curva, la specificità e la sensitività.

Per la **regressione lineare classica** rileggiamo le classi in termini di +-1 ai fini dell'implementazione del modello regressivo elementare. Otteniamo i seguenti risultati:

Accuratezza: 80%

Sensitività: 84.4%

Specificità: 76%

	Actual 1	**Actual 0**
Predicted 1	38	12
Predicted 0	7	38

Per la **regressione logistica** otteniamo i seguenti risultati:

Accuratezza: 78.95%

Sensitività: 80%

Specificità: 78%

	Actual 1	**Actual 0**
Predicted 1	36	11
Predicted 0	9	39

Per l'**analisi discriminante lineare** otteniamo i seguenti risultati:

Accuratezza: 81.05%

Sensitività: 86.7%

Specificità: 76%

	Actual 1	Actual 0
Predicted 1	39	12
Predicted 0	6	38

Per l'**analisi discriminante quadratica** otteniamo i seguenti risultati:

Accuratezza: 73.68%

Sensitività: 93.3%

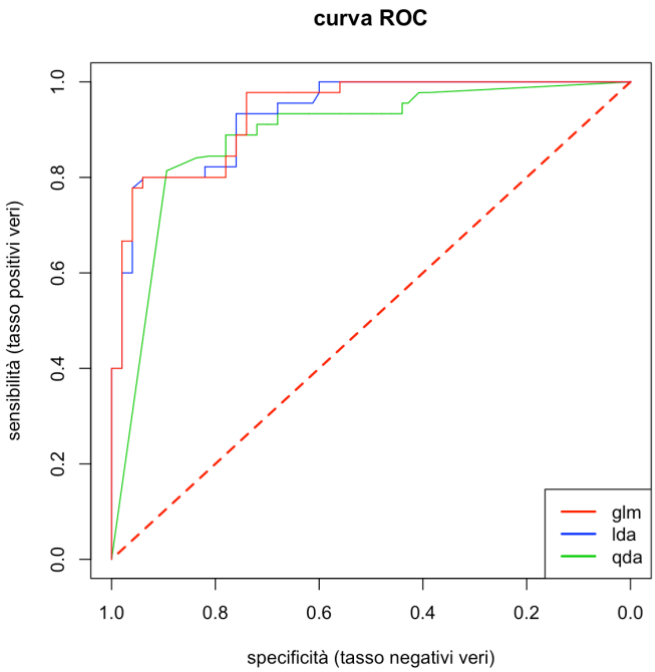
Specificità: 56%

	Actual 1	Actual 0
Predicted 1	42	22
Predicted 0	3	28

Tutti i classificatori presentano una sensitività maggiore della specificità.

In particolare, troviamo una netta differenza nell'analisi discriminante quadratica, di quasi 40 punti percentuali. Questo classificatore presenta inoltre la più scarsa accuratezza ottenuta in fase di training.

Gli altri tre classificatori presentano risultati abbastanza simili tra di loro in termini di accuratezza, sensitività e specificità.



	AUC
Regressione logistica	0.9356
Analisi discriminante lineare	0.9323
Analisi discriminante quadratica	0.8860

Confrontando le curve ROC e il valore dell'AUC per i tre classificatori sopra elencati, i migliori classificatori, secondo queste due metriche, risultano essere la regressione logistica e l'analisi discriminante lineare.

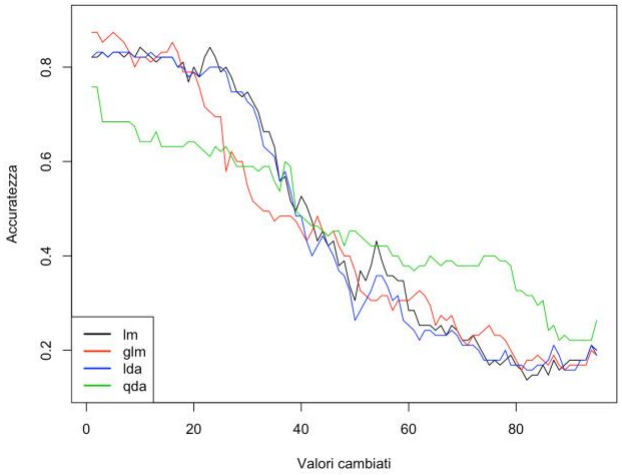
2.1. ROBUSTEZZA DEI METODI DI CLASSIFICAZIONE

Studiamo adesso la robustezza dei modelli scambiando l'etichetta di indici casuali, introducendo quindi informazioni false, e studiando l'andamento dell'accuratezza.

Per la regressione lineare classica lo scambio è avvenuto semplicemente cambiando di segno la classe (dopo averla riletta in termini di +-1), mentre per gli altri classificatori scambiando tra di loro 0 e 1.

Dall'analisi della robustezza dei modelli emergono i seguenti risultati.

Il modello di regressione lineare e l'analisi discriminante lineare presentano un andamento simile. La regressione logistica presenta l'andamento peggiore, perdendo una notevole percentuale di accuratezza già di fronte allo scambio di un valore compreso tra 20 e 30 etichette. Un comportamento molto distinto dagli altri classificatori è dato dall'analisi discriminante quadratica, che presenta un andamento quasi lineare.



3. PREDIZIONE E AUTOVALUTAZIONE

Verifichiamo la capacità di predizione dei modelli ripetendo più volte l'esperimento su campioni casuali così da avere un risultato statisticamente significativo.

Durante la fase di autovalutazione otteniamo i seguenti risultati in termini di accuratezza:

	media	**deviazione standard**
Regressione lineare	0.58	0.17
Regressione logistica	0.81	0.14
Analisi discriminante lineare	0.84	0.12
Analisi discriminante quadratica	0.68	0.12

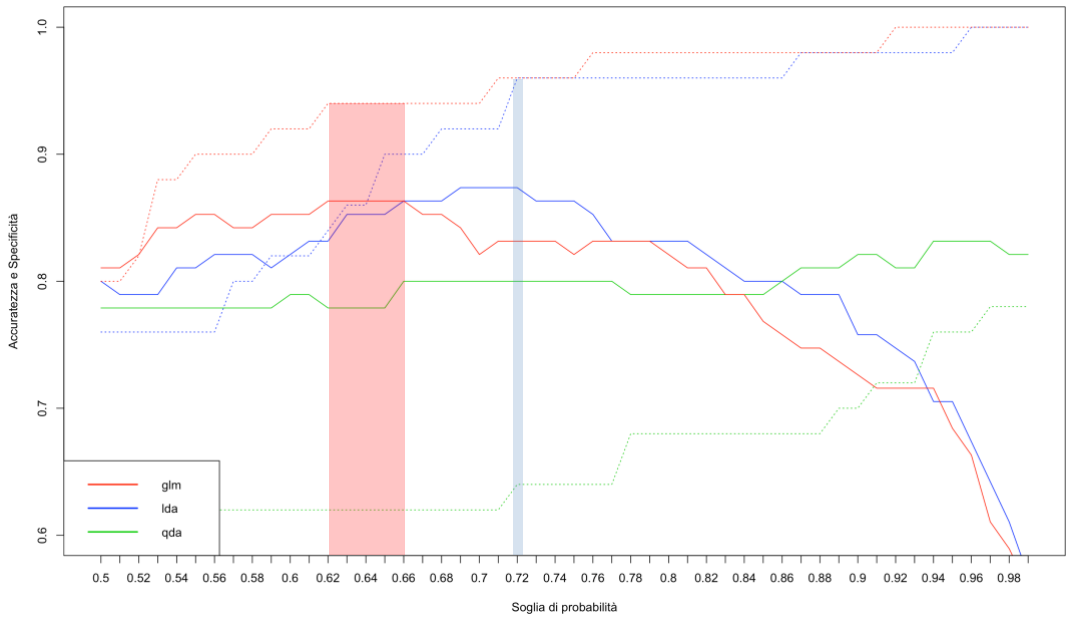
I risultati ottenuti in fase di autovalutazione non discostano molto da quelli ottenuti in fase di training per la regressione logistica, l'analisi discriminante lineare e l'analisi discriminante quadratica.

Riscontriamo un lieve aumento dell'accuratezza per la regressione logistica e l'analisi discriminante lineare. La regressione lineare classica presenta invece un'accuratezza media molto inferiore a quella ottenuta in fase di training con un valore di deviazione standard piuttosto elevato.

Questi risultati suggeriscono quindi che la regressione lineare possa non essere un metodo adatto per questa classificazione.

4. TRADE-OFF TRA ACCURATEZZA E SPECIFICITA’

Si è cercato di raggiungere un equilibrio tra specificità e accuratezza valutando questi due parametri per diversi valori della soglia di probabilità, aumentandola in modo da allargare l’intervallo di valori di probabilità per i quali un’osservazione viene considerata appartenente alla classe 0. Per quest’analisi sono stati considerati il modello di regressione logistica, l’analisi discriminante lineare e l’analisi discriminante quadratica.



E’ stato rappresentato con la linea tratteggiata l’andamento della specificità, mentre con la linea continua troviamo l’andamento dell’accuratezza.

Consideriamo adesso i migliori valori di trade-off (cercando di preservare il massimo valore di accuratezza rispetto alla specificità) per i classificatori che funzionano meglio, non considerando quindi l’analisi discriminante quadratica che presenta uno scarso valore di specificità.

Considerando l’analisi discriminante lineare, il miglior trade-off tra le due metriche può essere ottenuto con una soglia pari a 0.72, con la quale si raggiunge un’accuratezza circa pari all’87% e una specificità pari al 92%.

Considerando la regressione logistica, il miglior trade-off tra le due metriche può essere ottenuto con una soglia compresa all’incirca tra 0.62 e 0.66, con la quale si raggiunge un’accuratezza pari all’86% e una specificità pari al 94%.

CONCLUSIONI

Dai risultati ottenuti è emerso come l'analisi discriminante lineare e la regressione logistica siano i classificatori più adatti.

La regressione lineare classica presenta infatti scarsi risultati in fase di autovalutazione e l'analisi discriminante quadratica presenta valori di accuratezza, ma soprattutto di specificità, ridotti rispetto agli altri classificatori.

Tra i due migliori classificatori, nonostante il valore di specificità lievemente maggiore per la regressione logistica, è tuttavia preferibile l'analisi discriminante lineare per la sua maggiore robustezza e i suoi migliori risultati ottenuti in fase di autovalutazione.

Scegliendo quindi l'analisi discriminante lineare e fissando la soglia di probabilità pari a 0.72, otteniamo i seguenti risultati:

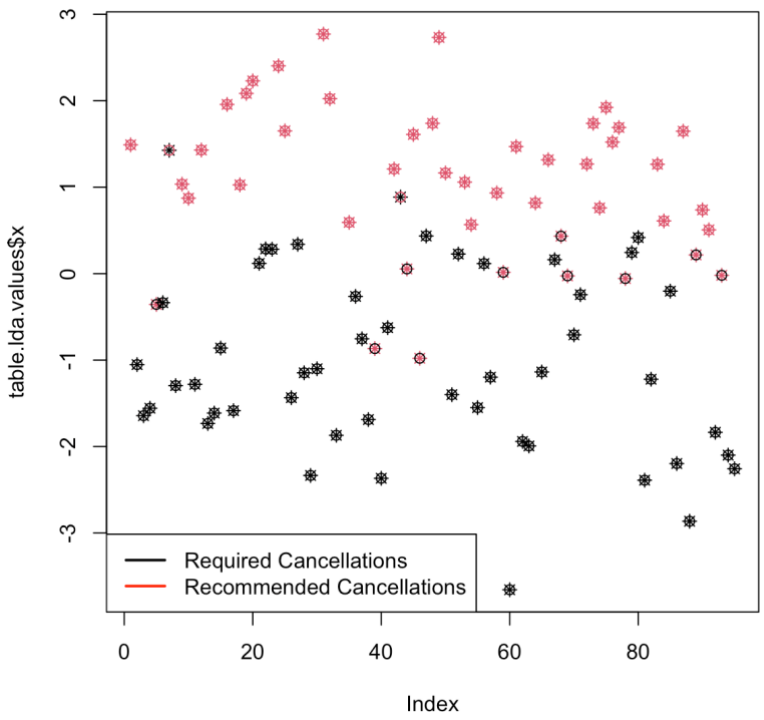
Accuratezza: 87.3%

Sensitività: 77%

Specificità: 92%

	Actual 1	**Actual 0**
Predicted 1	35	2
Predicted 0	10	48

Per questo classificatore e la soglia di probabilità scelta, abbiamo il seguente grafico che mostra il confronto tra le previsioni (rappresentate con cerchi) e le classi effettive (rappresentate al loro interno con stelle).



APPENDICE

L’operazione di pre-processing ha previsto il campionamento delle informazioni per Stato e la scelta di un sotto-insieme di fattori, escludendo quelli con un eccessivo numero di valori mancanti e quelli non rilevanti per la classificazione.
Sono stati utilizzati i dati relativi all’inizio di settembre, per i quali sono presenti un numero minore di osservazioni con valori mancanti.
Le righe di codice relative all’operazione di pre-processing sono presenti nel file R commentate, visto che è stata allegata la tabella contenente i dati in seguito all’operazione di pre-processing.