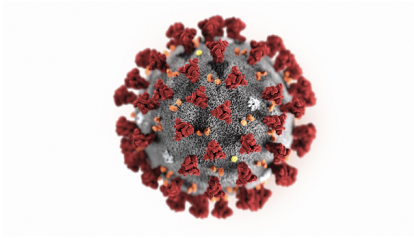
APRIL 2020

Covid-19 in Italia & Data Analysis



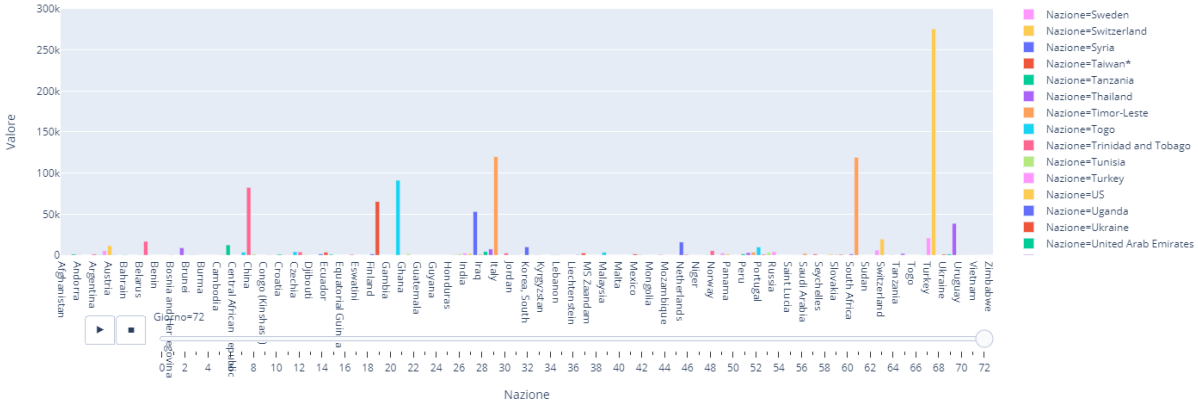
Abstract

*La **COVID-19** (più semplicemente conosciuto come **coronavirus** è una malattia infattiva respiratoria causata dal virus denominato SARS-CoV-2. I primi casi sono stati riscontrati a fine 2019 e attualmente questo virus si sta diffondendo rapidamente in tutto il mondo. L'Italia purtroppo è uno degli stati più colpiti da questa pandemia. Ho provato ad analizzare i dati italiani e confrontarli con il resto del mondo andando a produrre un modello finale di previsione della diffusione della pandemia. I dati che utilizzato a livello italiano sono forniti dalla protezione civile, mentre per le timeseries mondiali ho utilizzato i dati dell'Hopkins University.*

1 Data Analysis

Il **machine learning** è un metodo di analisi dei dati che automatizza la costruzione di modelli analiti. I dati sono dunque alla base. Lo scopo di questo mio esercizio è riuscire a capire se attraverso il machine learning è possibile creare uno o più modelli che descrivano e siano utili nelle previsioni di questa pandemia. Per creare un modello è sempre necessario fare uno buono studio dei dati e se possibile visualizzarli. Useremo degli strumenti quali pandas, plotly e sklearn durante questo studio (trovate il notebook completo a questo url). Ho suddiviso l'analisi dei dati in 4 categorie (è bene fare delle prove su colab per vedere i progressi in modo interattivo):

- **Dati Mondiali**:

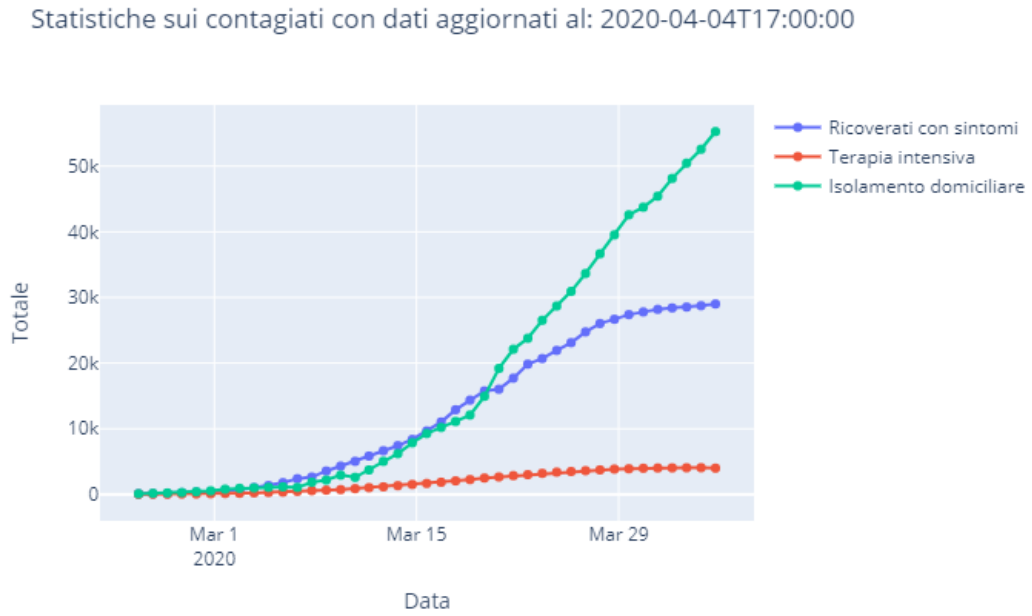


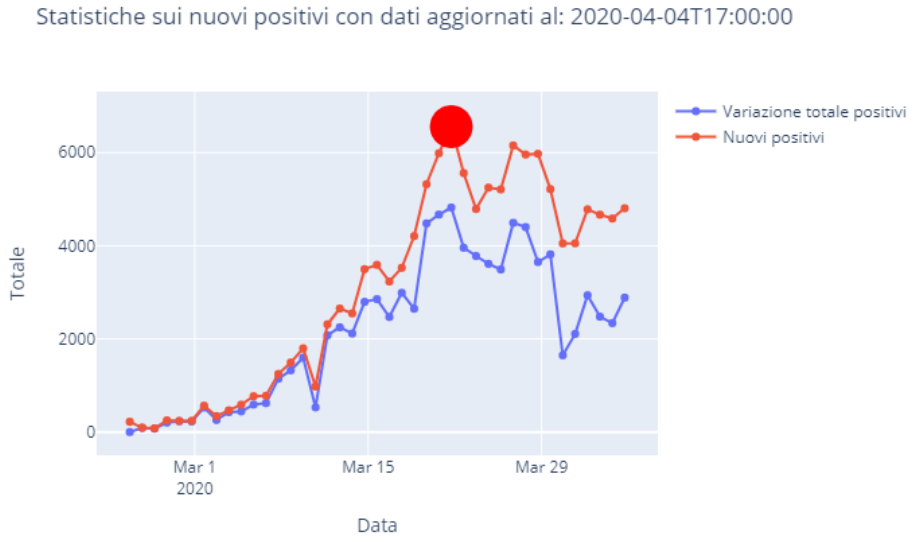
Sui dati mondiali ho realizzato un grafico animato utilizzando plotly che fa vedere l'evoluzione del virus nel mondo giorno per giorno. (Su ogni colonna è possibile ottenere una serie di informazioni diverse). Ho poi calcolato il tasso di mortalità per ogni nazione aggiornato all'ultimo report disponibile (purtroppo dai dati si evince che quello Italiano è uno dei più alti anche considerandolo e rapportandolo al numero di casi totali).

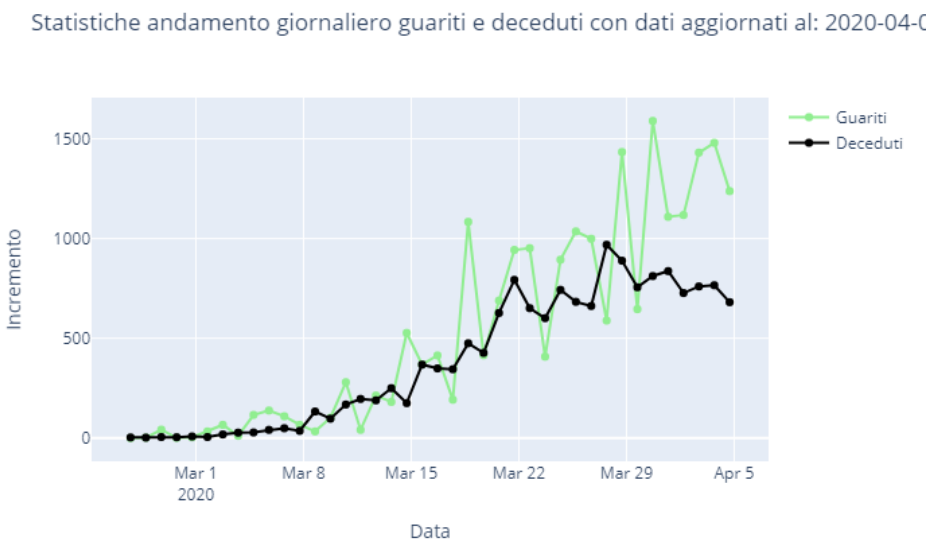
Bianca Francesco
francescobianca.github.io ☎ 3271046777 ✉ francescobianca96@gmail.com

- **Dati Nazionali**:

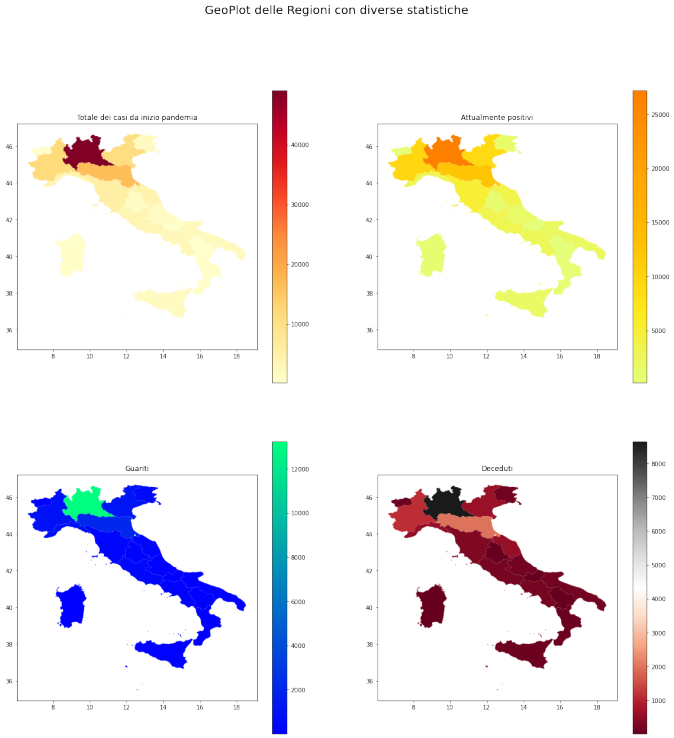
  I dati Nazionali sono quelli che probabilmente ognuno di noi guarda con maggiore interesse e che ci aiutano a comprendere quando supereremo il famoso picco. Negli ultimi giorni (sto scrivendo in data 5 Aprile 2020) le curve dei ricoverati in terapia intensiva e dei ricoverati con sintomi registrano valori in discesa e ciò lascia ben sperare. Un altro elemento importante è la variazione totale dei positivi quotidiana che fatica si sta stabilizzando ma fatica a scendere sotto i 2000 al giorno. In realtà ci sono molte variabili che vanno tenute in conto come l'estensione delle misure cautelari della Lombardia a tutta Italia, la chiusura di tutte le attività, il numero di tamponi effettuati ogni giorno ecc.



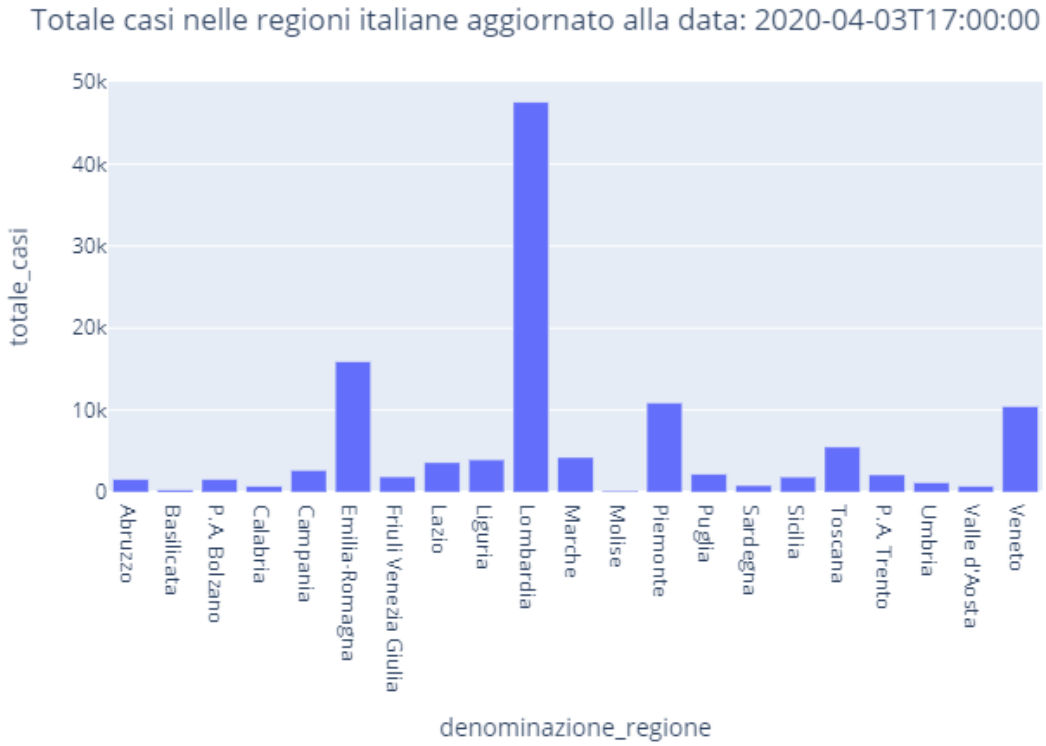




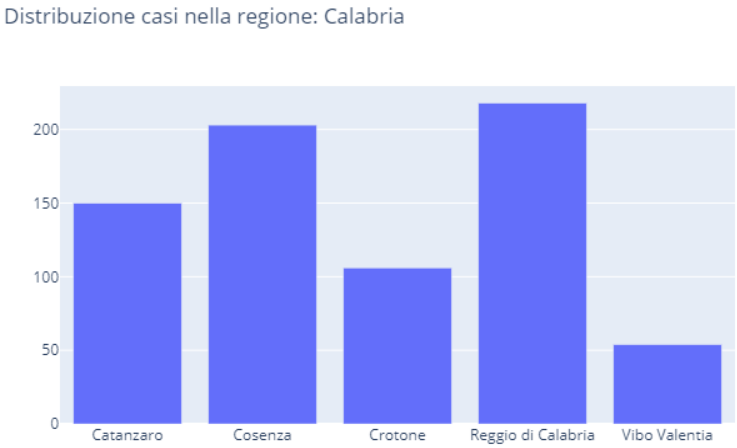
- **Dati Regionali**:



Le statistiche regionali seguono lo stesso schema dei dati nazionali e anche come si vede dalla figura sopra in tutte le statistiche la regione Lombardia ha il triste primato di regione più colpita.

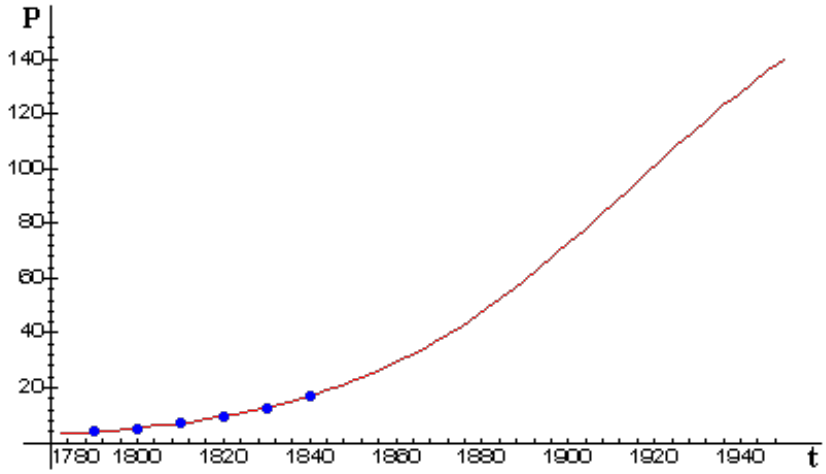


- **Dati Provinciali**:
  Sui dati provinciali non abbiamo molti elementi di analisi che ci vengono forniti dalla protezione civile italiana. Infatti ci viene fornito solamente il totale dei casi diviso per provincia. Riporto solo un piccolo grafico che riassume la distribuzione dei casi totali divisi per provincia nella mia regione.



2 Model Definition

Ho cercato di documentarmi su quali possibili algoritmi siano applicabili in queste tipologie di problemi. I dati da utilizzare per provare a creare un modello e fare una previsione sono la data e i casi totali relativi a quella data in modo da poter osservare la crescita della curva. In merito mi sono confrontato con un particolare problema presente nella biologia (chiamato **Exponential logistic growth**). In questo campo si studia vome crescono le popolazioni quando hanno risorse illimitate (e come i limiti delle risorse cambiano quel modello). In particolare mi sono focalizzato sulla *logistic growth*: nella crescita logistica, il tasso di crescita pro capite di una popolazione diventa sempre più piccolo man mano che la dimensione della popolazione si avvicina al massimo imposto da risorse limitate nell'ambiente, noto come capacità di carico *K*. Ritroviamo quindi dei pattern simili a quelli che si manifestano nella diffusione del *covid-19*. Un *logistic growth model* è chiamato anche *Verhulst model* in onore di *P. F. Verhulst* matematico belga che ha studiato questa idea nel 19esimo secolo applicandola alla crescita della popolazione degli Stati Uniti. Verhulst fece una previsione nel 1840 della popolazione degli Stati Uniti nel 1940 sbagliandosi di meno dell'1%. (approfondimento).

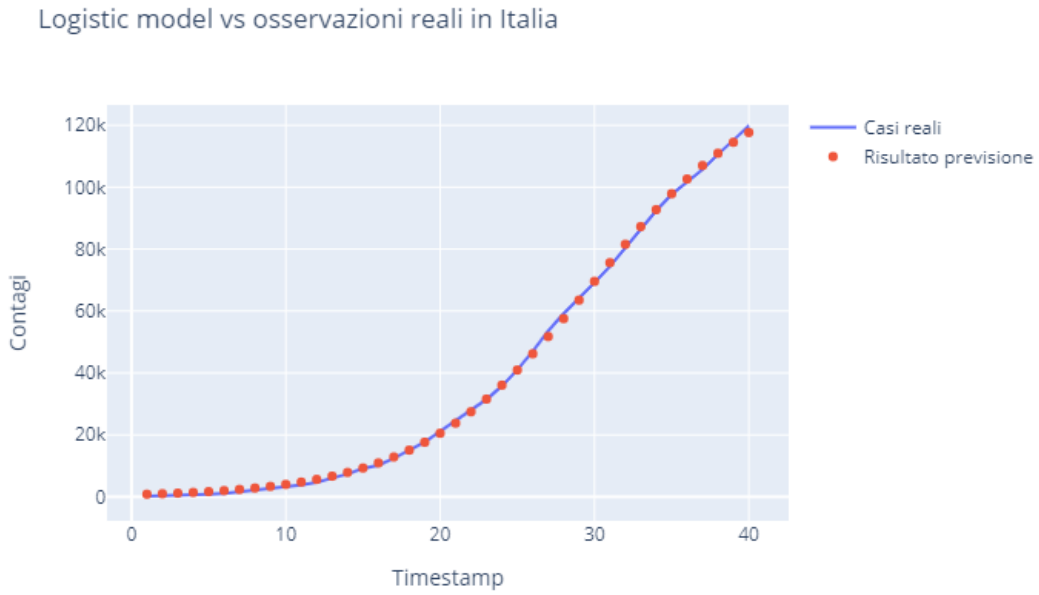


La formula del *logistic growth model* si può sintetizzare come:

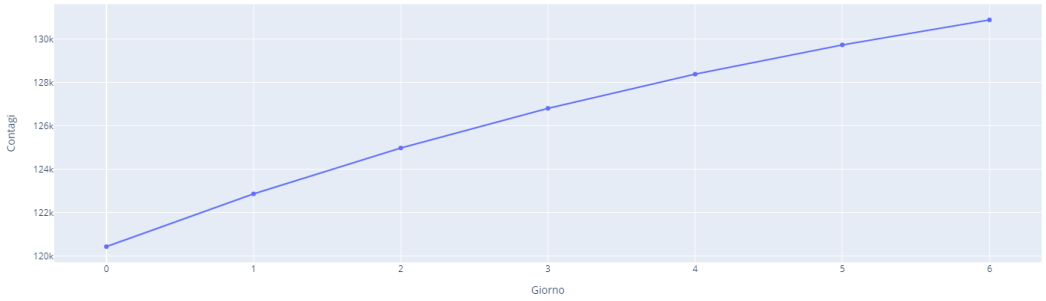
$$y(t) = \frac{c}{1 + a * e^{-bt}}$$

in cui:

- *y(t)* è il numero di casi in un dato tempo t.
- *c* è il valore limite, la capacità massima per y (numero infetti a fine epidemia)
- *b* deve essere maggiore di 0



Vediamo che la curva sembra adattarsi bene in fase di training. Naturalmente come nella maggior parte degli algoritmi di machine learning servono dati e quindi la qualità dell'algoritmo dobrebbe migliorare con l'aumentare dei dati. (Naturalmente voglio chiarire che il tutto è a fine didattico e quindi queste previsioni sono da considerare non veritiere). Proviamo a vedere una previsione che avevo effettuato qualche giorno fa che mostra come la curva dei casi totali va sempre di più fino ad appiattirsi (quando raggiunge valori stabili sarà da considerarsi un'ottima notizia).



Si potrebbe estendere questo lavoro provando altre tipologie di algoritmi epidemiologici. Vorrei come lavoro futuro provare ad implementare il SIR model (Model for Spread of Disease). La difficoltà sta nelle molte variabili che entrano in gioco quali ad esempio la considerazione del lockdown o i tamponi effettuati.