# Open data sulla pandemia di COVID-19 in Italia

Attraverso una ricerca nel web, si sono trovati alcuni open data che riportano i numeri dei positivi, dei ricoverati e dei morti in Italia durante la pandemia. La fonte principale di informazione è la Task force dedicata alla pandemia in seno alla Protezione Civile. Sui dati di questa elaboreremo l’analisi seguente. Altre fonti, che confermano i dati, sono quelle dell’Istituto Superiore di Sanità e del Coronavirus Resource Center della John Hopkins University. Si preferisce la fonte della Protezione Civile perché rispetto all’ISS fornisce open data con maggior grado di libertà (**☆☆☆** vs **☆**) e per la minor entropia rispetto ai dati del CRC, pure classificati a tre stelle, che però presentano l’intera serie storica per ogni stato del mondo, suddivisi in diversi file rispettivamente per i positivi, i ricoverati e i morti.

Dalla Protezione civile si può ottenere un file .csv che descrive l’andamento della pandemia in Italia: la keyword è un campo chiamato “1”, che identifica univocamente i 1076 record, ma è univoco anche il campo data di rilevazione; seguono poi i campi attributo come lo stato, il numero di ricoverati con sintomi, di pazienti in terapia intensiva, il totale degli ospedalizzati, la quantità di persone in isolamento domiciliare, il totale dei positivi, la variazione totale dei positivi, i nuovi positivi, i dimessi guariti, i deceduti, i casi da sospetto diagnostico, i casi da screening, il totale dei casi, i tamponi, i casi testati, un campo note ed altri campi di controllo e di dettagli sulle tipologie di testing. L’informazione che cerchiamo va trovata all’ultima riga, poiché i dati sono cumulativi e quindi le ultime date rappresentano i dati più aggiornati sull’andamento totale della pandemia. Questo non vale per il numero dei ricoverati, che invece richiede dei file di supporto dalla stessa fonte o delle operazioni sui dati raccolti.

Al 2 febbraio 2023 i positivi cumulati ammontano a 25488166 (al campo totale casi), i morti a 187272. Quanto ai ricoverati si possono calcolare o ricavare dalle altre fonti che riportano dati omologhi: questi ammontano a 4128664. Si deve notare che nel caso di positivi e ricoverati il dato conteggia più volte i casi di reinfezione.

È stata condotta un’indagine nella classe per determinare il numero di studenti che sono stati contagiati ed eventualmente ricoverati durante la pandemia. I risultati sono stati organizzati in un file .csv in cui un record header identifica una keyword ID e a seguire gli attributi Informazioni cronologiche (che registrano quando il record è stato aggiunto), Nome, Cognome, Città, Positivo, Ricoverato, Giorni Degenza. Da un primo confronto si può notare come la classe non sia un campione rappresentativo della prevalenza del virus nella popolazione né dei tassi di ospedalizzazione: probabilmente per l’età media significativamente inferiore la classe, pur avendo un tasso di diffusione del virus del 60%, di diciassette punti percentuali maggiore rispetto al dato nazionale, ha un tasso di ospedalizzazione nullo.

In analisi di questo tipo possono collaborare diverse figure professionali. Non è il caso delle fonti a cui abbiamo attinto, ma alle stesse si sono rifatte molte testate giornalistiche che hanno quotidianamente monitorato l’evoluzione della pandemia attraverso i dati, probabilmente avvalendosi di data journalist. Il lavoro massiccio di raccolta dei dati dalle singole strutture ospedaliere e dai presidi locali fino ai centri regionali di gestione della Sanità, che sono confluiti in seno alle task force governative deputate alla redazione dei bollettini sono un esempio di data engineering in cui diversi flussi costanti di informazione sono stati gestiti per tenere aggiornato un database centrale. I comitati scientifici, sulla base di questi dati, si sono avvalsi di data analyst per valutare le strategie di contenimento in essere, le differenze territoriali e demografiche ed eventuali modifiche dei piani d’azione. Dovendosi armonizzare il diritto alla salute e le libertà individuali, molti governi hanno cercato di limitare nel tempo e nei modi le restrizioni, tentando di prevedere con dei modelli matematici l’evoluzione della pandemia; lo stesso è stato fatto nelle accademie e in prestigiosi istituti di statistica ed economia: questo lavoro è un tipico esempio di data science.

La gravità della situazione ha imposto un buon grado di serietà scientifica in questi studi, per cui in tutte le fonti che abbiamo trovato si è associato un metadato comune: la nota metodologica sulle tecniche di raccolta dati e sui modelli di previsione. Inoltre, dalla fonte principale di questa analisi si può ricavare un altro metadato, che registra il numero di accessi al dato stesso: questo dà la misura della diffusione dell’informazione.