Esercizio - D3

Autore: Andrea C.

Data: 07.06.2023

Obiettivo: Ricercare e analizzare cinque fonti dati da internet che descrivano la situazione Covid al 2023.

A picture containing text, screenshot, font, document

Description automatically generated

Data Insight: L’analisi è stata svolta su cinque Dataset. Per semplicità, e vista la disuguaglianza di dati , sono stati presi come oggetto di ricerca i dati cumulativi dall’inizio della Pandemia 2020 fino al mese di maggio 2023. I database analizzati hanno tutti sicuramente le seguenti classi:

* Positivi Covid
* Dimessi/ Guariti
* Deceduti
* Ricoverati (laddove possibile)

PARTE 1

1. Fonte: Kaggle - [Corona virus latest data 2023 | Kaggle](https://www.kaggle.com/datasets/chitrakumari25/corona-virus-latest-data-2023) –

Licenza: ODC – PDDL

1. Data/Ora Consultazione: 06.06.2023 – 22.28
2. Stato del Servizio: Attivo
3. Descrizione della Fonte: interfaccia API (comunicano sistemi diversi con files .csv, .xls e .py) dati scaricabili, possibilità di filtrare prima di scaricare, Ente privato. Autore ha collezionato tutti i dati usando scrapy framework. – Aggiornamento dati settimanale.
4. Campi Rilevati:
5. *Casi Covid Totali*: Numero cumulativo di quante persone hanno avuto esito positivo al tampone semplice.
6. *Casi Covid Positivi Attuali*: Casi Covid Totali – Decessi accertati – Dimessi/Guariti
7. *Dimessi/ Guariti*: Questa statistica potrebbe essere imperfetta, perché ci potrebbero essere dei data reporting missing, incompleti e/o basati su differenti definizioni – o semplicemente calcolati in modo diverso nei vari DB utilizzati per estrapolare il dato finale. Inoltre, ogni paese ha un sistema diverso di misurazione dati.
8. *Deceduti* = numero cumulativo di decessi (tra i casi rilevati).
9. Tamponi effettuati
10. Popolazione/Paese
11. Modello 5 Stars: ««««« Le cinque stelle vengono attribuite poiché i file sono scaricabili in formato csv e ci sono anche dei collegamenti ad altri DB , quindi sono dei Linked Opened Data (LOD). Inoltre, il DB è raggiungibile tramite URL.
12. Fonte: Ministero della Salute link → [COVID-19 Italia - Desktop (arcgis.com)](https://opendatamds.maps.arcgis.com/apps/dashboards/0f1c9a02467b45a7b4ca12d8ba296596)

Licenza: Attribution 4.0 International (CC-BY-4.0)

Metadati: Link → [RNDT - Scheda Metadati (geodati.gov.it)](https://geodati.gov.it/geoportale/visualizzazione-metadati/scheda-metadati/?uuid=PCM%3ACOVID-19%3A05032020%3A093000) – Meta Tag → presenti nel codice

1. Data/Ora Consultazione: 06.06.2023/ 22.31
2. Stato del Servizio: Attivo
3. Descrizione della fonte: interfaccia API presente (come comunicano sistemi diversi con file .csv, .xls e .py), dati scaricabili, Tabelle + dashboards altamente interattive, con possibilità di filtrare prima di scaricare. Ente privato, Settore Sanitario.
4. Campi rilevati:
5. *Casi Totali Covid*: Numero cumulativo di quante persone hanno avuto esito positivo al tampone. semplice.
6. *Casi Covid Positivi Attuali*: Casi Covid Totali – Decessi accertati – Dimessi/Guariti.
7. *Dimessi/Guariti*: Questa statistica potrebbe essere imperfetta, perché ci potrebbero essere dei data reporting missing, incompleti e/o basati su differenti definizioni – o semplicemente calcolati in modo diverso nei vari DB utilizzati per estrapolare il dato finale.
8. *Deceduti* = numero cumulativo di decessi (tra i casi rilevati)
9. Modello 5 Stars: ««««« Le cinque stelle vengono attribuite poiché i file sono scaricabili in formato csv e ci sono anche dei collegamenti ad altri DB; quindi sono dei Linked Opened Data (LOD).
10. Fonte: Il Sole 24 Ore Link → [Coronavirus in Italia, i dati e la mappa (ilsole24ore.com)](https://lab24.ilsole24ore.com/coronavirus/?refresh_ce)

Licenza: - Attribution 4.0 International Public License.

Metadati: Dati presi e analizzati dal sito della protezione civile- I dati relativi all’indice di trasmissibilità dal sito Istituto Sanitario Superiore (Iss) e i dati internazionali dal sito ourworldindata. Link → [GitHub - pcm-dpc/COVID-19: COVID-19 Italia - Monitoraggio situazione](https://github.com/pcm-dpc/COVID-19)

1. Data/Ora Consultazione: 06.06.2023 ore 23.00
2. Stato del Servizio: Attivo – aggiornato al giorno 02.06.2023.
3. Descrizione della fonte: interfaccia API presente (come comunicano sistemi diversi con file .csv, .xls e .py), dati scaricabili, possibilità di filtrare direttamente tramite utilizzo di grafici interattivi. Focus su dati sia italiani che internazionali, un grafico specifico con il trend giorno per giorno (con dati giornalieri, cumulati, media mobile a 7 giorni). Un grafico scatter plot che indica la crescita nuovi casi su base settimanale. Un confronto anno su anno, la variazione percentuale giornaliera e per finire una rappresentazione grafica di tamponi, varianti virus etc. Quotidiano economico-politico-finanziario italiano.
4. Campi rilevati:
5. *Casi Totali Covid*: Numero cumulativo di quante persone hanno avuto esito positivo al tampone semplice.
6. *Casi Covid Positivi Attuali*: Casi Covid Totali – Decessi accertati – Dimessi/Guariti.
7. *Dimessi/Guariti*: Questa statistica potrebbe essere imperfetta, perché ci potrebbero essere dei data reporting missing, incompleti e/o basati su differenti definizioni – o semplicemente calcolati in modo diverso nei vari DB utilizzati per estrapolare il dato finale.
8. *Deceduti* = Numero cumulativo di decessi (tra i casi rilevati)
9. *Terapia Intensiva:* Sia valore totale che dato settimanale aggregato compresa una variazione % rispetto al totale della settimana precedente*.*
10. *Ricoverati con sintomi:* Sia valore totale che dato settimanale aggregato compresa una variazione % rispetto al totale della settimana precedente*.*
11. Isolamento Domiciliare = Sia valore totale che dato settimanale aggregato compresa una variazione % rispetto al totale della settimana precedente*.*
12. Modello 5 Stars: ««««« Le cinque stelle vengono attribuite poiché i file sono scaricabili, ci sono anche dei collegamenti ad altri DB quindi sono dei Linked Opened Data (LOD)– i dati per la costruzione dei grafici sono forniti dalla protezione civile, i dati sulla letalità e indice di trasmissione RT sono dati da Istituto Superiore di Sanità (ISS) rintracciabile tramite URL.
13. Fonte: La Repubblica in collaborazione con Gruppo Editoriale GEDI Link : [Coronavirus: la mappa aggiornata dei contagi in Italia, divisi per regioni e province (gedidigital.it)](https://lab.gedidigital.it/gedi-visual/2020/coronavirus-i-contagi-in-italia/)

Licenza: - Attribution 4.0 International Public License.

Metadati: I dati per elaborazione report sono presi sia dal sito del Ministero della Salute & Agenzia Nazionale per i Servizi Sanitari Regionali ( Agenas.gov.it); Meta Tag → presenti nel codice.

1. Data/Ora Consultazione: 07.06.2023 17:45
2. Stato del Servizio: Attivo – Pubblicazione dati settimanale (aggiornato al 26.05.23) , Aggiornamento dati ogni Venerdì.
3. Descrizione della Fonte: interfaccia API presente (come comunicano sistemi diversi .csv, .xls e .py), dati apparentemente non scaricabili, Tabelle e grafici di chiara e semplice visualizzazione ma statiche – non è possibile effettuare filtri/query di ricerca. C’è un grafico specifico per i contagi in Italia diviso per regione – e una specifica di nuovi casi settimanali per ogni 100k abitanti. Quotidiano Italiano appartenente a Gedi Gruppo Editoriale (Exor).
4. Campi rilevati:
5. *Casi Totali Covid*: Numero cumulativo di quante persone hanno avuto esito positivo al tampone semplice
6. *Casi Covid Positivi Attuali*: Casi Covid Totali – Decessi accertati – Dimessi/Guariti
7. *Guariti/Dimessi*: Questa statistica potrebbe essere imperfetta, perché ci potrebbero essere dei data reporting missing, incompleti e/o basati su differenti definizioni – o semplicemente calcolati in modo diverso nei vari DB utilizzati per estrapolare il dato finale.
8. D*eceduti* = numero cumulativo di decessi (tra i casi rilevati).
9. Ricoverati con Sintomi Covid.
10. Terapia Intensiva.
11. Isolamento Domiciliare.
12. Modello 5 Stars: ««««« Le cinque stelle vengono attribuite poiché i file sono scaricabili, ci sono anche dei collegamenti ad altri DB, quindi, sono dei Linked Opened Data (LOD).
13. Fonte: Worldometers link [COVID - Coronavirus Statistics - Worldometer (worldometers.info)](https://www.worldometers.info/coronavirus/#countries)

Licenza: Attribution 4.0 International

Metadati: utilizza trusted thousand sources in real time – questi dati sono anche utilizzati dai Governi ed enti i.e. UK Government -Financial Times – BBC etc. ; Meta Tag → presenti nel codice

Data/Ora Consultazione: 07.06.2023 – 18:01

1. Stato del Servizio: Attivo – Il Covid tracker è passato da una modalità live a una rilevazione giornaliera in data 01.02.2023
2. Descrizione della fonte: interfaccia API presente (come comunicano sistemi diversi .csv, .xls e .py), dati apparentemente non scaricabili, c’è in primo piano i dati rilevanti inerenti agli attributi Totale Casi (Mondo), Deceduti e Guariti/Dismessi dall’ospedale a livello mondiale. Nella seconda parte della pagina web è possibile utilizzare i dati strutturati, in formato tabellare, per filtrare in base al paese oggetto di ricerca. Ente privato composto da un team internazionale di sviluppatori, ricercatori e volontari con l’obiettivo di rendere disponibili dati e statistiche rilevanti ad un pubblico esteso.
3. Campi rilevati:
4. *Casi Totali Covid*: Numero cumulativo di quante persone hanno avuto esito positivo al tampone semplice
5. *Casi Covid Positivi Attuali*: Casi Covid Totali – Decessi accertati – Dimessi/Guariti
6. *Guariti/Dimessi*: Questa statistica potrebbe essere imperfetta, perché ci potrebbero essere dei data reporting missing, incompleti e/o basati su differenti definizioni – o semplicemente calcolati in modo diverso nei vari DB utilizzati per estrapolare il dato finale.

.

1. *Deceduti* = numero cumulativo di decessi (tra i casi rilevati).
2. *Paese di riferimento*
3. Casi nei passati 7 giorni
4. Variazione % settimanale Casi
5. Casi nei passati 7 giorni/ 1 M popolazione
6. Decessi negli ultimi 7 giorni
7. Variazione % dei decessi settimanali
8. Popolazione totale del Paese
9. Modello 5 Stars : «««« Le quattro stelle sono state attribuite perché è un dato aperto, rintracciabile tramite URL, non ho inserito la quinta stella solo perché non ho trovato i link che rimandano ai DB utilizzati ai fini della ricerca.

Definizione Tabella per Confrontare i Dati:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| FONTE | # CASI TOTALI | # DIMESSI/GUARITI | # DECEDUTI | # NUOVI CASI |
| Kaggle | 25.788.387 | 25.468.617 | 189.738 | 130.032 |
| Ministero Salute | 25.870.833 | 25.560.511 | 190.517 | 119.805 |
| Il Sole 24 Ore | 25.870.833 | 25.560.511 | 190.517 | 119.805 |
| La Repubblica | 25.857.572 | 25.545.397 | 190.392 | 121.783 |
| Worldometers | 25.870.833 | 25.560.511 | 190.517 | 119.805 |

* I Dati sono cumulativi e si riferiscono al Paese Italia

Definizione Campi/ Attributi Tabella:

* Casi Totali = numero cumulativo di quante persone hanno avuto esito positivo al tampone semplice
* Dimessi Guariti = tutti Coloro che hanno avuto il Covid ( ospedalizzati e non) e sono guariti.
* Deceduti
* Positivi attuali = coloro che attualmente risultano positivi ad un tampone Covid.

PARTE 2

Chiave Primaria

La chiave primaria per quanto riguarda la tabella nella parte 1 potrebbe essere la Fonte perché è quell’attributo che in maniera univoca mi identifica la riga - cioè è l’elemento univoco che non si ripete all’interno del DB.

Le Figure nel mondo dei Dati

### DATA ENGINEER:

La figura del Data Engineer si occupa della creazione della pipeline del dato stesso. È la prima figura che interviene sul dato, si occupa di svolgere una prima pulizia del dato.

Nel caso dei DB analizzati il D.E. si occupa della raccolta del dato raw in particolare del Back-end ( la logica che permette di capire i processi).Responsabile inoltre, della raccolta e memorizzazione dei dati stessi in maniera logica e che ne permette l’utilizzo ai fini di analisi.

### DATA ANALYST:

La figura che interviene subito dopo nel processo di raccolta trasformazione e analisi dei Dati.

* Prende i dati di nuovo, pulisce il DB, colleziona metadata e inizia a fare delle analisi.
* Analisi che devono essere in linea con un Business Problem specifico.
* Comunica con il D.E.
* I dati vanno quindi trasformati in base all’obiettivo da raggiungere

### Data Journalist

La figura combina dati da fonti diverse ed utilizzando infografiche riesce a fornire un prodotto facile ed intuitivo da visualizzare, anche per un utente non esperto nel campo di ricerca. Negli esempi nelle fonti La Repubblica e Sole 24 ore, sono state utilizzate, possibilmente, un Data Engineer per la raccolta e la logica di strutturazione dei dati, Un Data Analyst che ha raccolto e trasformato i dati per fornire delle soluzioni riguardo il problema iniziale ( Andamento Covid in Italia fino al 2023) e il Data Journalist magari in aiuto con il Data Analyst, ha creato delle infografiche user-friendly per l’utente finale.

### Data Scientist

La figura del D.S. per i DB interrogati personalmente penso non sia presente - in quanto sono stati ricercati e analizzati DB inerenti a serie storiche/ attuali e non modelli o DB che utilizzano modelli predittivi per fare un forecasting dell’andamento dell’epidemia futuro.

Metadati

In questo report sono stati ricercati, collezionati e analizzati cinque diversi Database. Vista l’eterogeneità di tali DB, sono stati rilevati sia metadati descrittivi (per esplicare meglio l’utilizzo del DB) e dove sono stati presi i dati poi analizzati.

Metadati amministrativi e gestionali (MAG ) per quanto concerne i diritti di accesso al documento – laddove presenti sono state inserite le licenze.

E Metadati strutturali – in quasi tutti i DB analizzati è stato possibile visualizzare (tramite Github) la definizione della struttura interna dei documenti e le eventuali relazioni con altri documenti.