**ESERCIZIO 3**

**COVID**

1. RECUPERARE 5 FONTI DATI DA INTERNET CHE DESCRIVANO LA SITUAZIONE COVID AL 2023 (CI DEVONO ESSERE ALMENO I CAMPI “POSITIVI COVID”, “RICOVERATI”, “MORTI”)
2. IN UN DOCUMENTO TESTO INDICARE LA FONTE, BREVE DESCRIZIONE DEL CONTENUTO, DESCRIZIONE DEI CAMPI E LA CLASSIFICAZIONE 5 STELLE DELLE FONTI SCELTE
3. DEFINIRE UNA TABELLA PER CONFRONTARE I DATI “NUMERO POSITIVI” COVID, “PERSONE RICOVERATE”, “CITTA’”, “GIORNI POSITIVI” DI OGNI PARTECIPANTE
4. DARE UN NOME ALLE ENTITA’ E AGLI ATTRIBUTI (CAMPI) IN MODO DESCRITTIVO

**COVID – PARTE 2**

* 1. DEFINIRE LA CHIAVE PRIMARIA PER OGNI TABELLA
  2. CONFRONTARE I DATI CLASSE/ITALIA E DESCRIVERE LA SITUAZIONE
  3. DESCRIVERE IN QUALE FASE INTERVIENE OGNUNA DELLE FIGURE STUDIATE (Data Scientist, Data Engineer, Data Analyst e Data Journalist) SE PRESENTI
  4. QUALI METADATI POTETE IDENTIFICARE IN QUESTA ESERCITAZIONE?

SVOLGIMENTO PARTE 1

Data: 11/06/2023

Fonti:

1. <https://www.salute.gov.it/imgs/C_17_monitoraggi_177_0_fileNazionale.pdf>
2. <https://lab.gedidigital.it/gedi-visual/2020/coronavirus-i-contagi-in-italia/mondo.php>
3. <https://data.humdata.org/dataset/coronavirus-covid-19-cases-and-deaths?force_layout=desktop>
4. <https://covid19-country-overviews.ecdc.europa.eu/key_epidemiological_indicators.html>
5. <https://ourworldindata.org/covid-cases>

descrizione fonti:

1. La prima fonte è un file pdf del Ministero della Salute aggiornato al 24 maggio 2023.

In questo documento ci sono molti dati in formato grafico e tabellare, i campi sono vari come: incidenza settimanale di nuovi casi, tasso di ospedalizzazione settimanale, tasso di mortalità ecc. Ma nonostante ciò secondo la classificazione a 5 stelle di Tim Berners Lee questo dato viene valutato con una stella siccome è in formato pdf e quindi è possibile solo scaricarlo e leggerlo ma non si può elaborarlo.

1. la seconda fonte è un sito web aggiornato al 10 marzo 2023, dove ci sono principalmente due campi: nuovi contagi e decessi. Entrambi i campi non hanno numeri precisi ma sono in percentuale per ogni milione di abitanti. Si possono trovare tre tipi di grafici interattivi: grafico a barre, grafico a linee e grafico a mappa. Secondo la classificazione a 5 stelle anche questi sono dati valutati con una stella visto che non sono né scaricabili né si possono elaborare.
2. La terza fonte è il sito web World Health Organization dove è possibile scaricare un file in formato csv, i campi sono: data del dato riportato, paese, codice paese, stato, nuovi casi, totale casi, nuovi morti, totale morti.

Questi dati essendo aperti, strutturati ed elaborabili sono classificati 3 stelle.

1. La quarta fonte è un sito web dove ci sono diversi grafici e c’è anche un csv scaricabile. Ci sono molti campi come: casi, morti, ospedalizzazioni, test, terapia intensiva. La classificazione è di 3 stelle visto che c’è un file scaricabile in formato csv, per quanto riguarda i grafici invece è di una stella in quanto sono solo consultabili.
2. La quinta fonte è tratta dal sito OurWorldInData dove il campo principale sono i nuovi casi aggiornati quotidianamente. I Campi sono sia in numero preciso sia in percentuale ad un milione di abitanti Anche questo dataset, per il file scaricabile in csv è di 3 stelle.

Confrontando i dati si intuisce che sono contrastanti tra loro ma che comunque la situazione Covid-19 in Italia negli ultimi tempi sta andando a migliorare.

Il data engineer è colui che ha raccolto i dati e li ha salvati, nei dataset visibili a noi utenti non risalta la sua figura poiché ha un ruolo di backend. Il data scientist e data scientist in questi casi non hanno un ruolo fondamentale visto che l’obiettivo di questi dati era puramente informativo. Se ci fosse come obiettivo capire i motivi attraverso questi dati della diminuzione di casi positivi e di morti allora in quel caso entrerebbe in gioco il data analyst, se invece l’obiettivo fosse capire attraverso quei dati cosa succederà in futuro e consigliare modi di fare per andare a migliorare allora entra in gioco il data scientist. Il data journalism invece potrebbe essere colui che ha esposto i dati attraverso i siti web.

I metadati che si possono identificare nelle varie fonti sono ad esempio:

* 1. “tasso di morti per ogni milione di abitanti”
  2. “Tasso di occupazione dei posti letto in Area Medica da pazienti con COVID-19 per giorno del monitoraggio settimanale”
  3. “La mappa si colora in base alla media dei nuovi decessi su 1 milione di abitanti”

Nella terza fonte c’è addirittura una sezione denominata “metadati” dove ci indica la fonte, il collaboratore, periodo di riferimento dei dati, l’ultimo aggiornamento, visibilità, licenza ecc…

Vincenzo Manzo