Title

Overview: Under considered dataset is a collection of the COVID-19 data maintained by *Our World in Data*. They update it daily and will keep updating throughout the duration of the COVID-19 pandemic. It includes the data of 65 different variables. The Dataset contains information from 22 February 2020 to till now. It consists of 132644 records with 65 different columns such as total_cases, new_cases, total_deaths etc. First, we analyze the dataset, pre-process, and apply machine learning algorithm. Finally, evaluate the model and calculate the qualitative as well as quantitative results.

Projective Objectives: Analyze the COVID-19 dataset to explore meaningful information and train a machine learning model so that to predict new cases and new death at early stage. Government will be beneficial from this project to take suitable action at early stage after prediction.

Dataset Description: This is the general overview of the dataset, as it indicate that there are 65 unique column, and 132644 total records. There are 3848452 cell which contains NaN values that mean there is need to pre-process the data. There is no duplicate records etc.

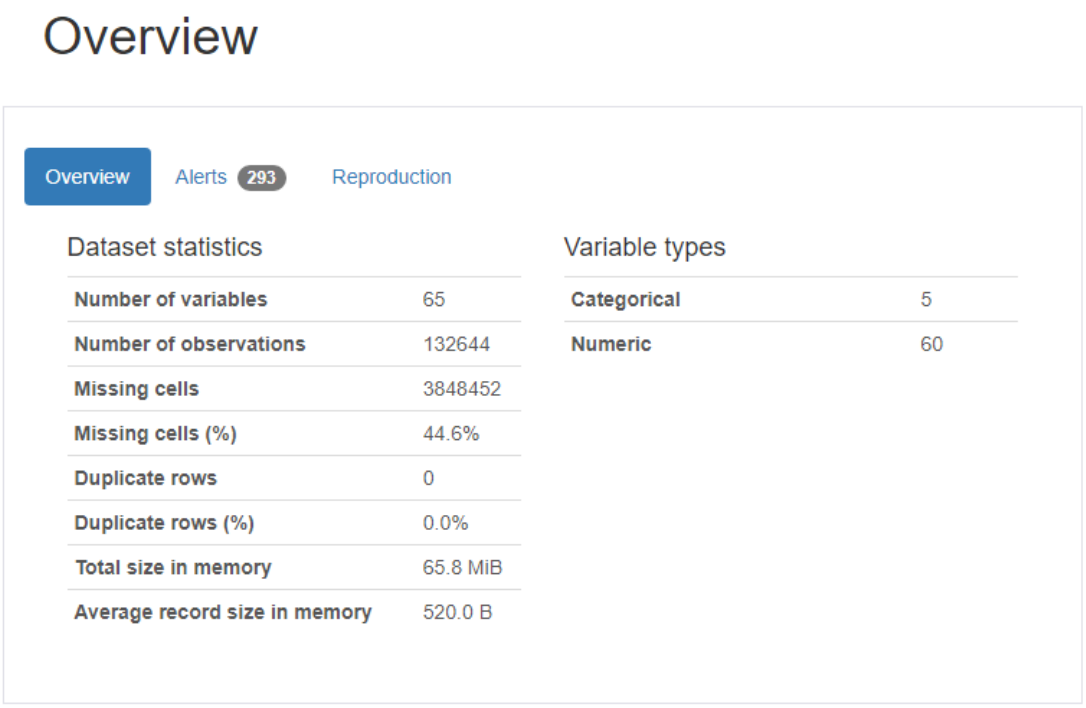

Figure 1: Dataset Description

Data Preprocessing:

- Check the duplicate values in dataset and remove.
- Unified the datatype of each column.
- Calculate NaN values in each column and replace with value 0.
- Check the negative values in columns and apply appropriate action.
- Split the dataset in to 80% and 20% ratio for training and testing the algorithm.

Results:

- Dataset contains 65 columns and 132644 records
- There are 44.0% cells with missing values
- Several columns contain NaN values
- Some columns contain negative values which is human error
- New case and death rate of age more than 65 is comparatively high
- K Nearest Neighbor algorithm can efficiently predict the new cases and new death before occurrence
- Government can take the appropriate decision before occurrence

Limitations:

- Dataset does not contain the information about recovered cases
- Dataset contains several columns which don't make sense for this project
- Several columns contain NaN values
- New_deaths columns contain negative values which mean there is human error which should be considered
- Due to very large dataset, machine learning algorithms can't train accurately

Tools:

- **Numpy:** a library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.
- **Pandas:** a library offers data structures and operations for manipulating numerical tables and time series.
- **Pandas_Profiling:** an open source Python library with which we can quickly do an exploratory data analysis with just a few lines of code.
- **Matplotlib:** a plotting library for the Python programming language and its numerical mathematics extension NumPy
- **Sklearn:** Scikit-learn is a free software machine learning library for the Python programming language.