Capstone Project Report

Samanta Rana, #0810971, Data Analytics, Capstone project

St. Clair College, Mississauga

Objective:

"Unveiling COVID-19's Impact: Data Analytics on Death and Vaccination Datasets."

Introduction

Title:

Unveiling the Insights of the Covid Dataset:
A Comprehensive Analysis of Covid Deaths and Vaccinations

The Covid-19 pandemic has left an indelible mark on humanity, challenging societies worldwide and transforming the way we live and interact. As this unprecedented crisis unfolded, researchers and policymakers sought to harness the power of data to understand its complexities better and devise effective strategies to combat the virus. In this context, our Capstone project takes on the monumental task of excavating and analyzing the vast Covid dataset, with a keen focus on two paramount aspects: Covid Deaths and Vaccinations.

The dataset at the heart of our study stands as a testament to the magnitude of the pandemic's impact, encompassing an impressive 10,868,428 instances. Each instance represents a unique snapshot of time, capturing various attributes that detail the evolution of the pandemic from its inception in January 2020 to the present day. The richness of this dataset empowers us to delve into a plethora of factors that have shaped the trajectory of Covid-19, allowing us to draw meaningful correlations and glean profound insights.

Crucial to our investigation are the two pivotal dimensions: Covid Deaths and Covid Vaccinations. The former focuses on the sobering reality of lives lost during the pandemic. By meticulously analyzing the CovidDeaths subset, we aim to unearth patterns and trends that shed light on the mortality rate under different circumstances. Our exploration encompasses critical phases of the pandemic, from its early emergence to the aftermath of vaccination campaigns. Understanding how the death rate has evolved over time and across regions is essential to comprehend the virus's behavior and adapt response strategies accordingly.

Furthermore, our analysis delves into the demographic aspect of Covid Deaths, probing how age distribution influences mortality rates. This knowledge can inform targeted measures to protect vulnerable age groups. Moreover, we investigate the socio-economic impact on mortality, as disparities in healthcare access and resources may have affected outcomes for different economic strata.

The second vital dimension, Covid Vaccinations, is central to understanding our path to recovery. Within the CovidVaccinations subset, we explore the efficacy of vaccination efforts in curbing the pandemic's spread. Delving into the time taken by officials to develop and distribute vaccines provides invaluable insights into the magnitude of the challenge faced by researchers, policymakers, and healthcare providers in rapidly responding to the crisis.

An equally vital aspect of CovidVaccinations is the public's response to vaccination campaigns. We analyze data points that reflect the duration and efficacy of public awareness campaigns aimed at fostering vaccine acceptance. Understanding the dynamics of vaccine hesitancy and acceptance is crucial for future vaccination initiatives, ensuring that accurate information reaches the masses and fostering trust in vaccination as a vital tool in combating infectious diseases.

Additionally, our study recognizes the tireless efforts of healthcare professionals, hospitals, and researchers. Their role in spearheading vaccination drives, administering shots, and studying vaccine efficacy has been instrumental in the battle against Covid-19. We endeavor to highlight their contributions and acknowledge their sacrifices, which have been pivotal in safeguarding communities and saving lives.

In conclusion, our research project seeks to navigate the intricacies of the Covid dataset to glean profound insights into Covid Deaths and Vaccinations. By analyzing this wealth of information, we aim to contribute significantly to the body of knowledge surrounding the pandemic's impact and the efficacy of vaccination efforts. Our findings hold the potential to inform public health strategies, guide policy decisions, and foster resilience in the face of future health crises.

First phase CRISP - DM

Understanding the business problem :-

Here are some potential research goals:

1. **Predictive Modeling**:
    - **COVID-19 Spread Prediction**: Develop models to predict the spread of COVID-19 cases over time, considering various factors like vaccination rates, population density, and mobility.
2. **Vaccination Analysis**:
    - **Vaccine Efficacy**: Evaluate the effectiveness of different COVID-19 vaccines in preventing infection and severe outcomes.
    - **Vaccination Impact**: Assess the impact of vaccination campaigns on reducing case numbers, hospitalizations, and deaths.
3. **Epidemiological Studies**:
    - **Disease Trends**: Analyze trends and patterns in COVID-19 cases, deaths, and recoveries.
    - **Hotspot Identification**: Identify regions or areas with a higher risk of outbreaks.
4. **Healthcare Resource Allocation**:
    - **Hospital Capacity Analysis**: Study the availability and utilization of hospital facilities and resources, and make recommendations for resource allocation.
    - **Optimizing Healthcare Response**: Develop models to optimize the allocation of healthcare resources during a surge in cases.
5. **Public Policy and Interventions**:
    - **Impact of Interventions**: Evaluate the effectiveness of various public health measures and interventions (e.g., lockdowns, mask mandates, social distancing).
    - **Policy Recommendations**: Provide evidence-based recommendations for policymakers to manage and mitigate the impact of the pandemic.
6. **Demographic and Socioeconomic Analysis**:
    - **Vulnerability Analysis**: Identify demographic and socioeconomic factors that correlate with a higher risk of infection or poor outcomes.
    - **Equity and Access**: Assess disparities in vaccine distribution and healthcare access.
7. **Mutations and Variants**:
    - **Genomic Analysis**: Study the genetic mutations and variants of the virus and their implications for transmission and severity.
8. **Behavioral Insights**:
    - **Public Behavior Analysis**: Examine how public behavior and compliance with guidelines impact the spread of the virus.
9. **Surveillance and Early Warning Systems**:

- **Early Detection**: Develop models for early detection of potential outbreaks and emerging variants.
- **Surveillance and Monitoring**: Implement a system for continuous monitoring of COVID-19 data and trends.

10. **Vaccine Deployment Strategy**:
    - **Optimal Distribution**: Determine the optimal strategy for vaccine distribution, considering factors like population density, vulnerability, and vaccine availability.
11. **Educational Campaigns**:
    - **Assessing Education Impact**: Evaluate the impact of public health education campaigns on public behavior and vaccine acceptance.
12. **Long-Term Effects**:
    - **Study Long-Term Health Effects**: Investigate the potential long-term health consequences for individuals who have had COVID-19.

Data Mining processes

The first step while initiating analysing the data was to clean it thoroughly, because -

1) it is a live data and so, can contain many impurities.

2) during the initial phase of Covid, the entries which got registered were not uniform and timely.

3) the data is of varied measures; implying normalization is required.

Another aspect of this dataset is that it contains over 3 million entries, hence, the project has started from cleaning only the attributes required for analysis.

Post-preprocessing, once the dataset is cleaned and prepared for the analysis, the first and most generalized analyses is carried out on the total number of cases as against the total deaths occurred. The trend was varied when the nations from all around the globe are considered. Following is the review of the analysis of total number of cases versus total number of deaths.

]

Data Pre-processing :

The dataset in concern contains numerous null values, count is different for varied fields, and hence the handling of null values is also different for all the fields. Also, since the dataset contain over 3 million entries, with the scope of complexity and chances of loss of data, complete data at once is not cleaned rather the fields required as per the research questions is cleaned accordingly.

Four major processes involved in cleaning are :-

1) The whole data contained null data cells at the beginning, implying that the early data was not recorded and probably not registered due to the sudden outbreak of Covid. Hence, those entries were fille with '0' .

2) Then the whole data is filtered year and then month wise and 'rolling means' is employed to fill in the null values after observing the definite pattern in the dataset. Since, all columns have different pattern to them, so, rolling mean of window 4- 18 and for the column vaccinations, window of 50 was taken too. The priority here is kept for the definite pattern to be observed so that no wrong analysis can be done.

3) to eradicate the null values for the column = 'population', firstly the data is divided nation-wise and then the mean of the national population is used to fill the null values.

4) for filling up the missing age of the demographic from the column = 'age', the dataset is divided into sub-groups of age and then median is taken to fill up the null values of the age

Thank you