Data Mining Project Proposal: Post-COVID-19 Analysis on Healthcare

Shubh Sanjeevkumar Mehta
Department of Computer Science, Indiana University,Bloomington

Sarthak Choudhary
Department of Computer Science, Indiana University,Bloomington

Niveditha Bommanahally Parmeshwarappa
Department of Computer Science, Indiana University,Bloomington

project-nibomm-shumehta-sarchou

Abstract

As of August 2021, the World Health Organization (WHO) reported over 210 million confirmed cases of COVID-19 and more than 4.4 million deaths worldwide, though these figures likely underestimate the true extent of the pandemic.[1,2] This proposal aims to conduct comprehensive research and analysis on the post-COVID-19 healthcare landscape

Keywords

COVID-19, healthcare analysis, predictive modeling, time-series forecasting, pattern discovery

1 Introduction

The COVID-19 pandemic has had a significant impact on global health. Despite our ongoing battle against the virus, our understanding of the natural history, clinical course, and consequences of COVID-19 remains incomplete. Emerging evidence suggests that approximately 10-20 percentage of those affected experience a range of mid and long-term effects after their initial recovery, known as "post-COVID-19 condition."[4] This proposal aims to comprehensively research and analyze this post-COVID-19 healthcare landscape.

Previous Work

Despite our ongoing battle against the virus, our understanding of the natural history, clinical course, and consequences of COVID-19 remains incomplete (3).

COVID-19, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), not only inflicts physical health challenges but also has profound emotional consequences that are yet to be fully understood. The impact is particularly severe among individuals with chronic diseases such as heart disease, diabetes, cancer, chronic obstructive pulmonary disease, chronic

kidney disease, and obesity, which increase the risk of severe illness from COVID-19 [5,6]. Other factors like smoking and pregnancy also elevate the risk [7].

2 Methods

This proposal aims to conduct comprehensive research and analysis on the post-COVID-19 healthcare landscape. We have selected a rich dataset comprising over 600,000 records, covering various regions (countries), age groups, and underlying health conditions.with Objectives like Post-COVID Health Effects to Investigate and categorize the mid and long-term health effects of COVID-19 and analyze their prevalence among different age groups, regions, and pre-existing medical conditions. Alone with Geographical Variations ,Age-Related Effect vary across different age groups.

Dataset(link : https://catalog.data.gov/dataset/conditions-contributing-to-deaths-involving-coronavirus-disease-2019-covid-19-by-age-group)

The primary goals of this project are threefold:

2.1 Predictive Modeling

1. Data Preprocessing: Cleanse the data of any inconsistencies, handle missing values, and prepare it for analysis.
2. Feature Engineering: Develop new features that can potentially improve the model's performance, such as population density or mobility indices.
3. Model Selection: Experiment with various machine learning algorithms, including ensemble methods like Random Forest and Gradient Boosting, to predict mortality rates.
4. Model Evaluation: Use cross-validation and a hold-out test set to evaluate model performance, employing metrics such as RMSE and MAE for regression tasks.

2.2 Time-Series Forecasting

1. Trend Analysis: Decompose the time series to understand underlying trends, seasonality, and irregular components.
2. Forecasting Models: Implement time-series forecasting models such as ARIMA, SARIMA, and Prophet to predict future trends in COVID-19 mortality.
3. Model Validation: Validate forecasts using time-series cross-validation and assess accuracy using metrics like the Mean Absolute Percentage Error (MAPE).

2.3 Pattern Discovery

1. Clustering: Use unsupervised learning algorithms to identify clusters in the data that may represent different patterns of mortality.
2. Association Analysis: Apply algorithms like Apriori to discover associations between different conditions and COVID-19 deaths.
3. Visualization: Employ visualization tools to present the findings, such as heatmaps for correlation analysis and dendrograms for hierarchical clustering results.

3 Expected Outcomes

The project is expected to deliver:

1. A set of predictive models with the capability to forecast COVID-19 mortality rates with high accuracy.
2. A comprehensive time-series analysis report that outlines the forecasted trends for the upcoming period.
3. A pattern discovery analysis that uncovers the underlying structures and associations in the data.

References

1. WHO coronavirus (COVID-19) dashboard. Geneva: World Health Organization; 2021 (https://covid19.who.int/ accessed 31 August 2021).
2. COVID-19. World Health Statistics. Geneva: World Health Organization; 2021 (https://www.who.int/data/gho/publications/worldhealth-statistics, accessed 31 August 2021).
3. Long Covid: what is it, and what is needed? London: Royal Society; 23 October 2020. DES7217.
4. https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-(covid-19)-post-covid-19-condition
5. Rosenthal N, Cao Z, Gundrum J, Sianis J, Safo S. Risk factors associated with in-hospital mortality in a US national sample of patients with COVID-19. JAMA Netw Open 2020;3(12):e2029058. Erratum in: JAMA Netw Open 2021;1:e2036103[REMOVED IF= FIELD] CrossRefexternal icon
6. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature 2020;584(7821):430–6. CrossRefexternal icon
7. Centers for Disease Control and Prevention. People with certain medical conditions. Updated March 29, 2021. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html. Accessed April 8, 2021.