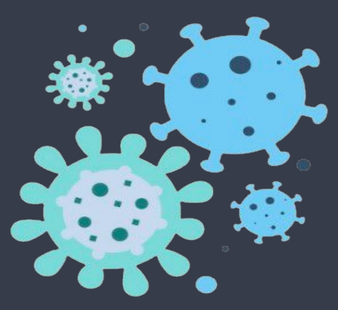


Covid19 - Severity and Mortality Analysis

Georgia Tech

CSE6242
Spring 2024

Saithanusri Avirneni (savirneni3), **Yoonha Kim** (ykim3056), **Harsha Ramesh** (hramesh8), **Maria Tariq** (mtariq37), **Thao Truong** (ttruong42)

1 Motivation and Introduction

- COVID is still responsible for many deaths worldwide, with over 6 million deaths in 2020 alone [1]. Despite the availability of vaccines, ongoing outbreaks fueled by mutant variants underscore the crucial need for real-time diagnostics.
- Our goal is to develop a Python predictive model using the Kaggle Mexican patient COVID-19 dataset, paired with interactive visualization in Tableau. We aim to enhance understanding of patient demographic and health conditions that lead to severe COVID-19 outcomes, enabling more efficient patient care.

Historically 20% of COVID hospitalized patients need ICU care [2] thus the model is crucial for healthcare workers, administrators, policymakers, and patients

- Advances Real-time Decision-making and Patient Care: By integrating comprehensive patient data analysis with real-time predictive capabilities, the model identifies high-risk COVID-19 patients promptly, potentially saving lives and improving management strategies.
- Innovates in Data Visualization and Informing Policy: Our use of Tableau for interactive visualization, combined with the detailed analysis of COVID-19 severity factors, enables more informed public health strategies and policy making, allowing better preparedness for COVID-19 spikes.

2 Approach

Algorithm

A Random Forest model was built in Python and used to predict patients at high risk of needing medical interventions based on underlying health data and age

- **Robustness and Efficacy**: The Random Forest classifier is renowned for its robustness and efficacy in handling complex, high-dimensional datasets for classification tasks.
- **Diverse Ensemble Construction**: During training, the classifier constructs multiple decision trees, with each tree contributing to the ensemble. Introducing randomness during tree construction enhances the diversity of the ensemble models, resulting in better predictions.
- **Versatile Feature Handling**: The Random Forest classifier excels in handling various data types and managing interactions between features and non-linear data relationships. This capability is particularly valuable in medical datasets where such complexities are common.

Building the final random forest algorithm involved three key phases:

Data Preparation → Development → Deployment

Data Preparation

- Remove missing values and non predictive columns such as medical unit and USMER number
- Create new features such as death flag from date died, age group from age, and risk level from Covid test results and patient information on hospital, ICU admission, intubation, and death

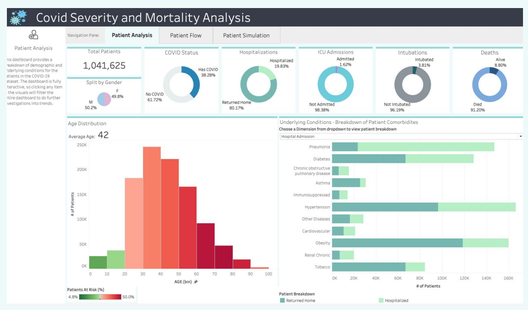
Development

- To improve our model, we refined the representation of influential features, particularly age. We enhanced our model by categorizing age into decades ('10s', '20s', '30s', etc.), addressing non-linear effects and reducing bias, thus improving interpretability.
- Model Optimization: Based on the importance scores, we streamlined the model by removing less impactful features. This not only simplified the model but also focused the training process on more significant predictors, improving the overall performance.

Model Deployment

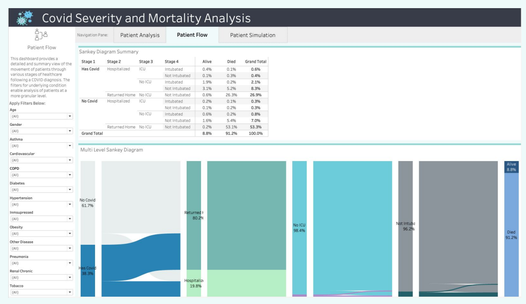
- The final trained random forest model (deployed via TabPY), along with two datasets created during model development, were used to create a complex interactive Tableau dashboard.

Interactive Visualization in Tableau



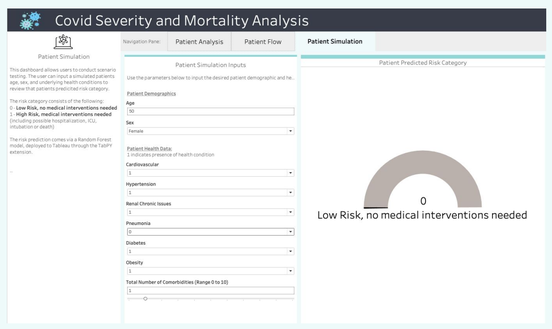
Patient Analysis

- Dashboard offers high-level values for patient medical interventions and dynamic breakdown of underlying conditions
- Clicking any item in the visuals filters the entire dashboard for deeper trend investigation
- Users explore overarching trends while delving into specifics



Patient Flow

- Detailed and summary view of patient movement post-COVID diagnosis
- Multilevel Sankey visualization illustrates intricate pathways of patient flow across different stages
- Filters for underlying conditions allow for more granular analysis of patients



Patient Simulation

- Users can perform scenario testing by creating simulated patients and viewing the predicted risk classification
- Powered by a trained Random Forest algorithm deployed to Tableau via the TabPY extension
- Tableau parameters serve as the arguments feeding into the script_real calculation

- The intuition behind our approaches lies in providing users with interactive visualizations that enable them to explore overarching trends while delving into specifics and patient scenario testing, empowering informed decisions and pattern identification in COVID-19 patient data.
- The novelty of the approach is the integration of multiple advanced visualization techniques and real-time predictive analytics into a user-friendly interface that can be implemented in healthcare settings to support COVID patient management.

3 Data

Data Source: Covid-19 patient data on Kaggle

How to obtain: directly accessible from Kaggle using the link: **Covid-19 Dataset**

Characteristics:

- Size on disk: 58 MB
- 21 unique features, 1,048,576 unique patient entries.
- Sourced from Mexican government, anonymized patient data.
- Patient hospitalization, intubation, ICU admission, and date of death records along with demographics (age, gender) and health conditions

4 Experiments and Results

- Utilized a split of 70% training and 30% testing data to rigorously assess model performance, measuring accuracy and precision across multiple models.

Model	Accuracy	Precision
Initial Random Forest Model	62%	63%
Improved Random Forest Model	66%	65%
Multilayer Perceptron Model (MLP)	71%	70%

- The Multilayer Perceptron (MLP) exhibited improved accuracy; however, the extended computation time required to run the complex model rendered the trade-off unfavorable for deployment into TabPY when dealing with a large dataset
- Testing various feature combinations lead to dropping 5 non influential features. Pneumonia, Age, Diabetes, Obesity, Hypertension were shown to be highly influential in predicting High Risk Covid patients and were used in the final model alongside other features.

5 Future Work

- Reliance on Mexican patient data limits universality but allows potential for global expansion or tailored models.
- Current model provides binary predictions, which may oversimplify outcomes. Future versions could estimate probability scores to assess risk more granularly. A probabilistic approach would enable more detailed risk assessments and predictions, crucial for decision-making processes in complex domains like healthcare.
- Continuous refinement and consideration of alternative approaches are vital for maximizing the model's utility in diverse healthcare settings.

6 References

1. M. Cascella, M. Rajnik, A. Aleem, S. C. Dulebohn, and R. Di Napoli, “Features, Evaluation, and Treatment of Coronavirus (COVID-19),” PubMed, 2024. doi: link
2. S. S. Zakariaee, N. Naderi, M. Ebrahimi, and H. Kazemi-Arpanahi, “Comparing machine learning algorithms to predict COVID-19 mortality using a dataset including chest computed tomography severity score data,” Scientific Reports, vol. 13, no. 1, p. 11343, Jul. 2023, doi: link