Machine Learning Models To Predict Covid Mortality

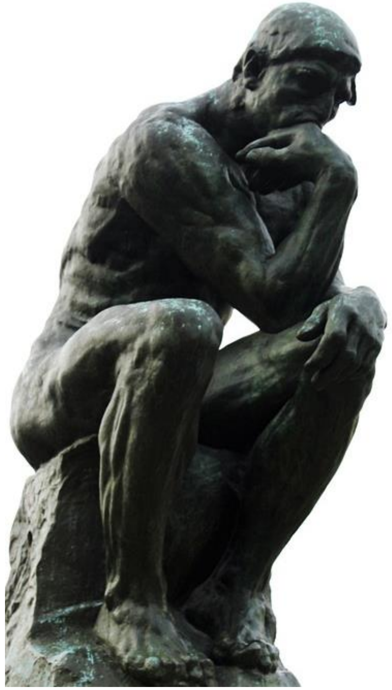
Rutgers Bootcamp for Data Science

Abraham Abate, Ashish Shukla,
Jialin Huang, Roger Vroom
March 23, 2023

Image: Corona Borealis Studio/Shutterstock.com

Team Introduction

Abraham Abate
Cruncher of numbers and dominator of databases



Ashish Shukla
Philosopher of "One more thing"

Jialin Huang
Artistic and Visual Director

Roger Vroom
[redacted]

Background

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people infected with COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment.

The main goal of this project is to build a machine learning model that, given a Covid-19 patient's current symptom, status, and medical history, will predict whether the patient is high risk or not.

Objectives

We decided to focus on Covid-19 due to its continued impact on the world's population.

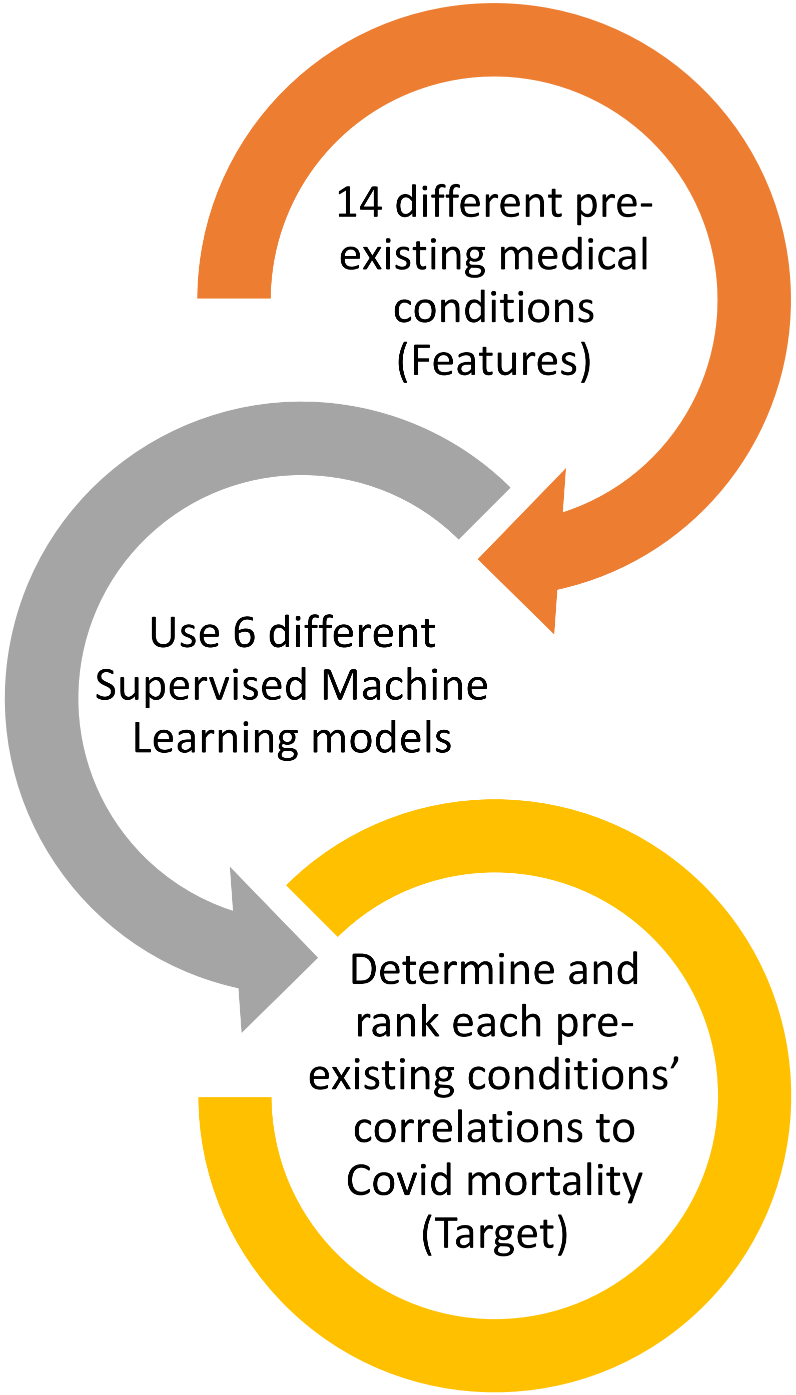
Dataset, provided by Mexican government, consists of 21 unique features and 1,048,576 unique patients

Compare and evaluate performances of different machine learning models

Predict a person’s likelihood contract severe Covid by pre-existing conditions (features)

Develop website to visualize medical conditions correlation to Covid and mortality

Data Exploration

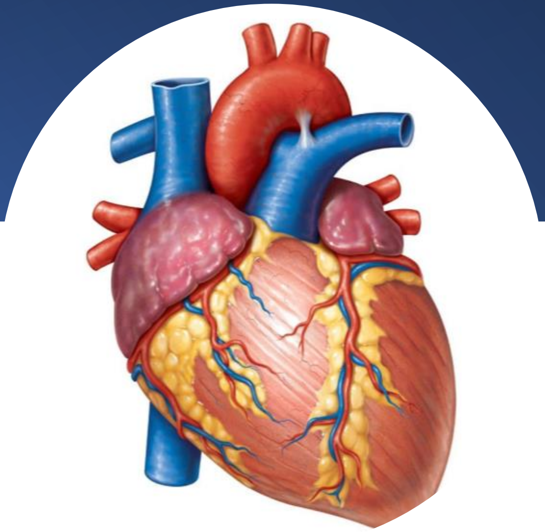


14 different pre-existing medical conditions (Features)

Use 6 different Supervised Machine Learning models

Determine and rank each pre-existing conditions’ correlations to Covid mortality (Target)

Pre-existing Diseases or Conditions

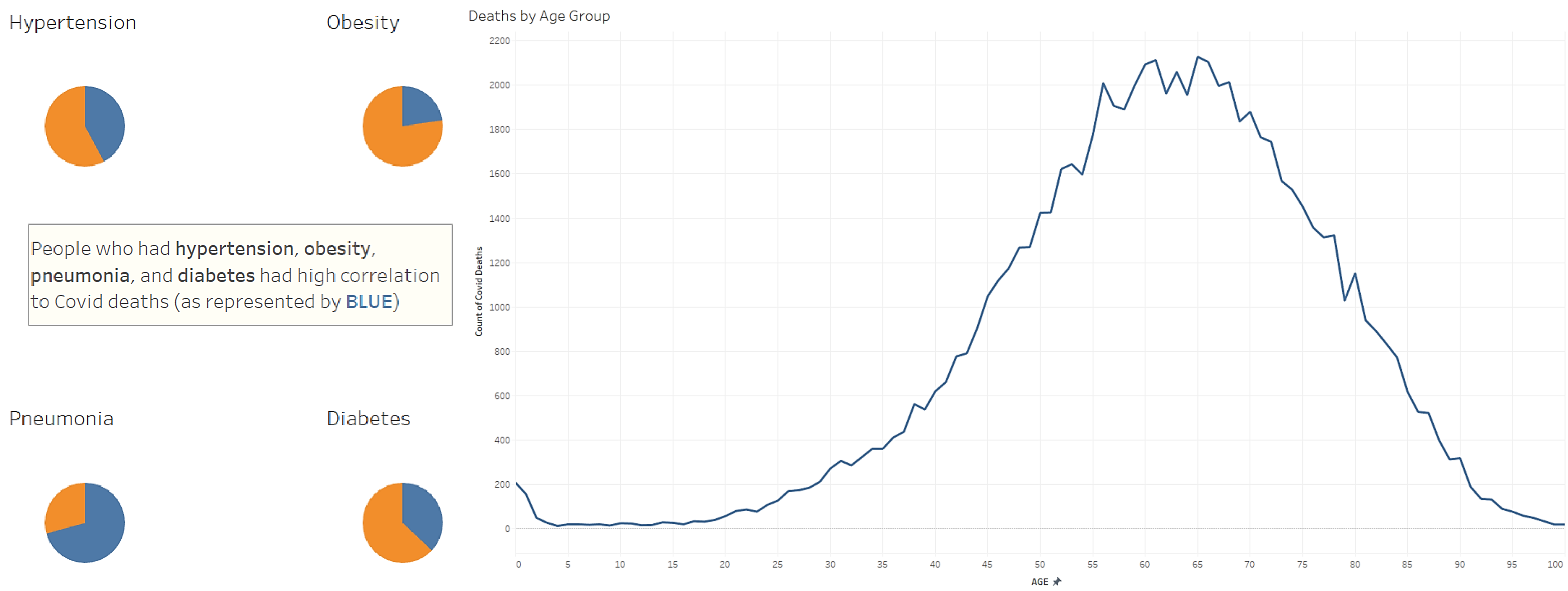


Older people, and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, and cancer, have higher chance to develop serious Covid symptoms, possibly leading to mortality

https://www.mayoclinic.org/diseases-conditions/coronavirus/in-depth/coronavirus-who-is-at-risk/art-20483301

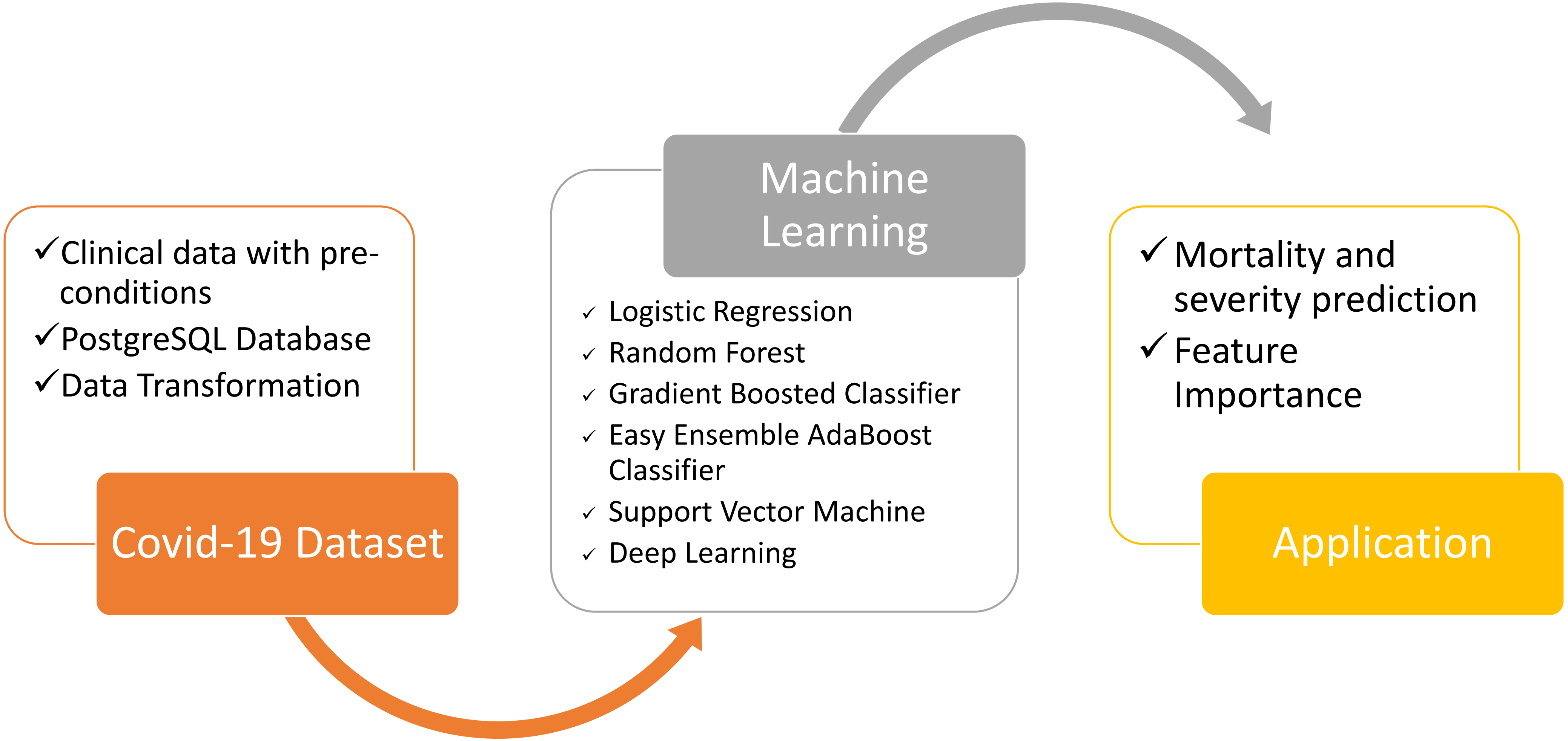
Tableau Visualization

to be demonstrated in website



People who had **hypertension**, **obesity**, **pneumonia**, and **diabetes** had high correlation to Covid deaths (as represented by **BLUE**)

Prediction Model/Machine Learning



Covid-19 Dataset

- ✓ Clinical data with pre-conditions
- ✓ PostgreSQL Database
- ✓ Data Transformation

Machine Learning

- ✓ Logistic Regression
- ✓ Random Forest
- ✓ Gradient Boosted Classifier
- ✓ Easy Ensemble AdaBoost Classifier
- ✓ Support Vector Machine
- ✓ Deep Learning

Application

- ✓ Mortality and severity prediction
- ✓ Feature Importance

PostgreSQL Database

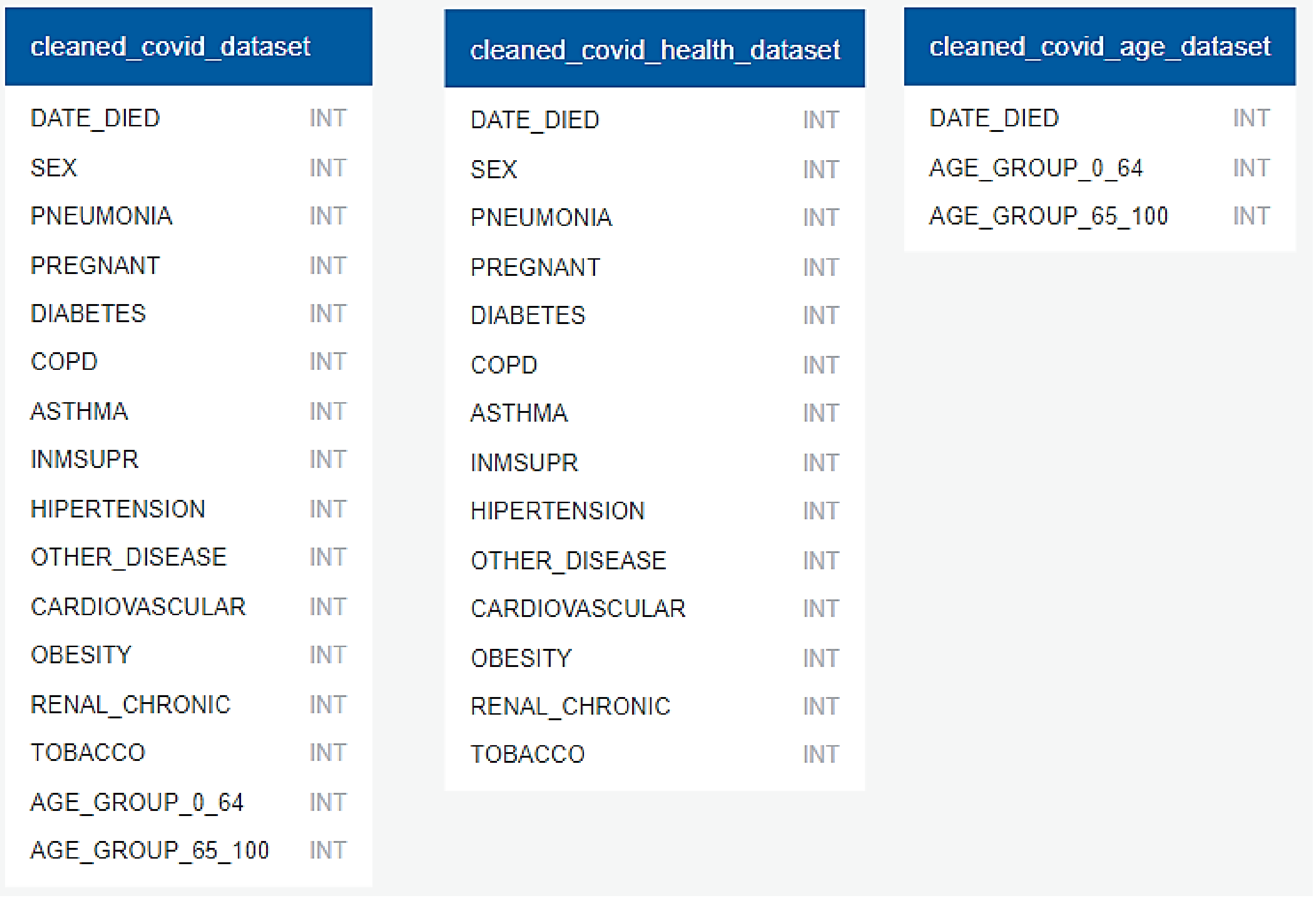
Process/ steps to clean up/ transform data

Features are in Boolean (1 for yes, 2 for no)

All missing and NA values were dropped (values >2)

DATE_DIED dropped “9999-99-99” indicating survival

Age group divided to **0-64** and **65-100**



cleaned_covid_dataset	
DATE_DIED	INT
SEX	INT
PNEUMONIA	INT
PREGNANT	INT
DIABETES	INT
COPD	INT
ASTHMA	INT
INMSUPR	INT
HIPERTENSION	INT
OTHER_DISEASE	INT
CARDIOVASCULAR	INT
OBESITY	INT
RENAL_CHRONIC	INT
TOBACCO	INT
AGE_GROUP_0_64	INT
AGE_GROUP_65_100	INT

cleaned_covid_health_dataset	
DATE_DIED	INT
SEX	INT
PNEUMONIA	INT
PREGNANT	INT
DIABETES	INT
COPD	INT
ASTHMA	INT
INMSUPR	INT
HIPERTENSION	INT
OTHER_DISEASE	INT
CARDIOVASCULAR	INT
OBESITY	INT
RENAL_CHRONIC	INT
TOBACCO	INT

cleaned_covid_age_dataset	
DATE_DIED	INT
AGE_GROUP_0_64	INT
AGE_GROUP_65_100	INT

Machine Learning Prediction

	Logistic Regression	Random Forest	Gradient Boosted Classifier	Easy Ensemble AdaBoost Classifier	Support Vector Machine	Deep Learning
Training Score	**0.914**	0.896	0.896	0.849	**0.913**	
Testing Score	**0.912**	**0.911**	**0.912**	0.845	**0.911**	**0.912**

Six ML models were trained on top of these 14 features to predict patients' mortality or discharge outcomes (target).

The logistic Regression model performs best with an accuracy of 91.4%, followed by SVM (91.3%) and Deep learning (91.2%).

The trained models were then tested on the test dataset. Three models (Logistic Regression, Gradient Boosting Classifier, neural network) had the best performance with an accuracy of 91.2%, followed by random forest and support vector machine (91.1%).

Confusion Matrix

	Logistic Regression	Random Forest	Gradient Boosted Classifier	Easy Ensemble AdaBoost Classifier	Support Vector Machine
Accuracy	0.912	0.912	0.911	0.896	0.911
Precision	0.61	0.61	0.59	0.39	0.61
Recall/ Sensitivity	0.42	0.39	0.46	0.84	0.39
F1	0.50	0.48	0.52	0.53	0.48

Precision rate of all ML models are at ~ 0.60 and recall rate at ~ 0.40 except Easy Ensemble

Easy Ensemble has the highest recall/sensitivity rate at 0.84, but the lowest precision rate at 0.39

F1 scores of all ML models are consistent at ~ 0.50

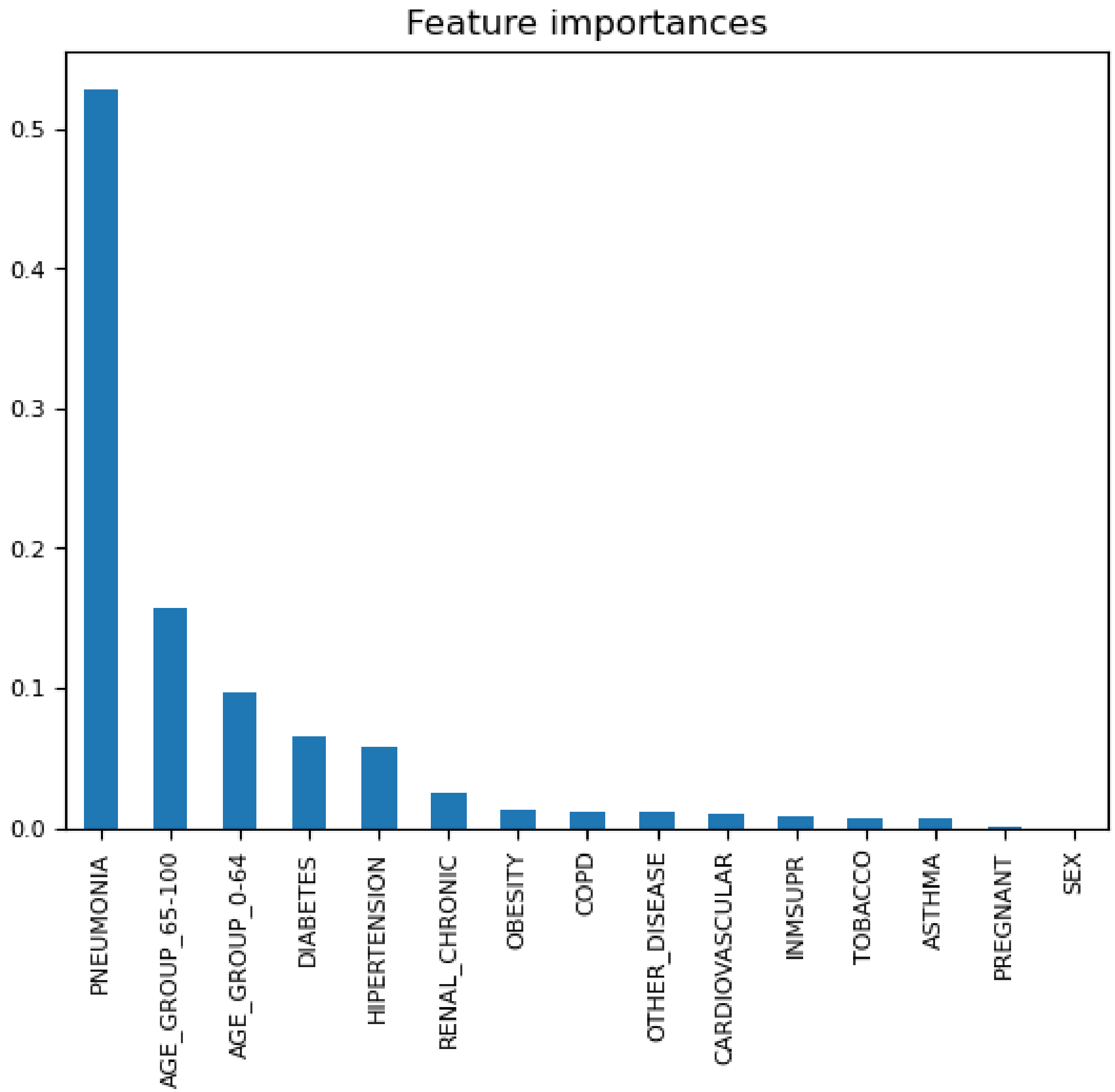
Obstacles/Challenges

Only 10% accuracy score for SVM

- Resolved by changing y-prediction scaled data to y-prediction data

CDC data was too large (>90 MB) including 95 normalized features

Feature Importance for Random Forest model



Features were ranked from the highest (most useful to predict target) to the lowest (least useful, “noise”)

Pneumonia and age were the two top features and contributed the most to the accuracy of the model

Additional Resources

Tableau:
https://public.tableau.com/app/profile/jialin.huang3459/viz/CovidmortalitygroupprojectMar-2023/Story1

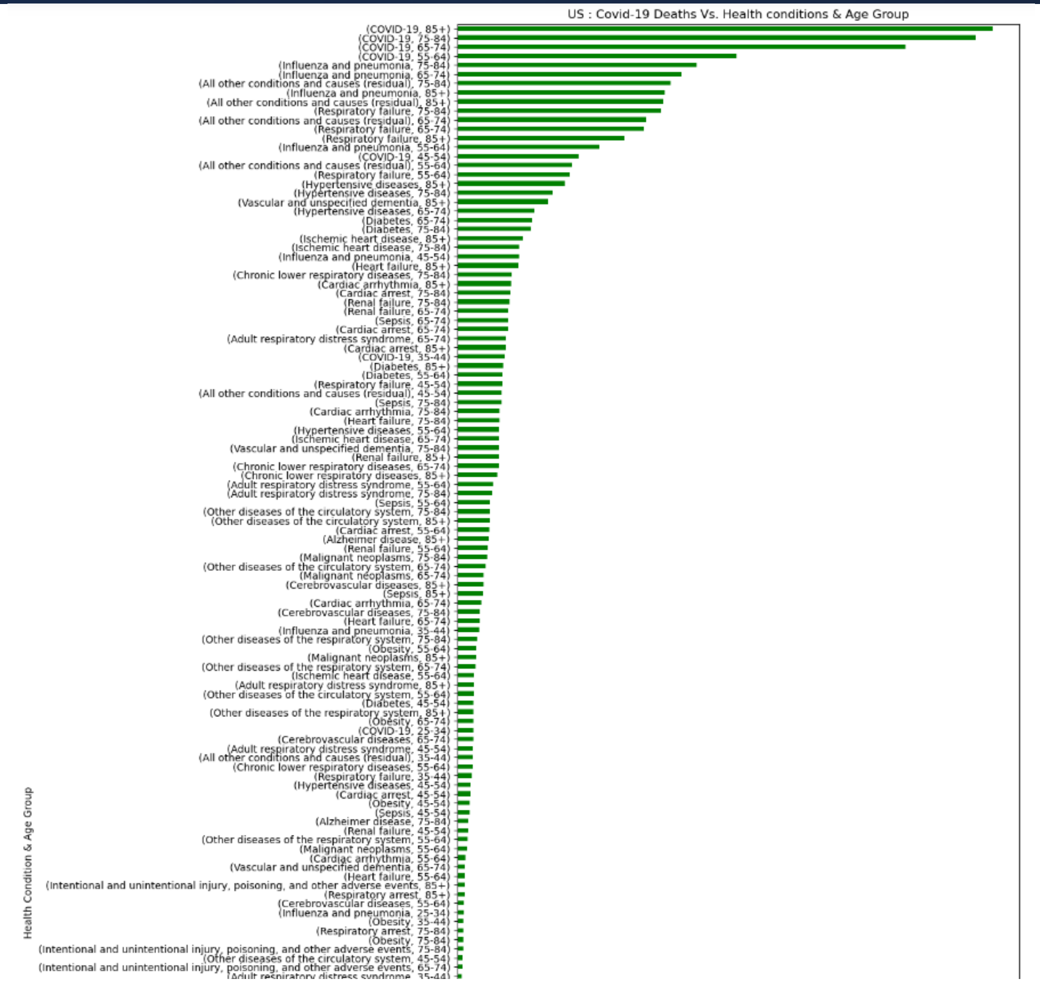
Github:

https://github.com/rvroomiii/group_hub

Website:

https://rvroomiii.github.io/group_hub/

Future Analysis



Recommendations for future analysis:

- US CDC data use in conjunction (state to state comparison, urban vs rural)
- US CDC data (>90 MB, 95 normalized features) too large for machine learning/model prediction

Anything that the team would have done differently:

- Further break down age group into more subgroups
- Compare datasets from different sources
- Website in development to predict user's severity with Covid based on medical conditions and age

Credits

Data Source:
https://www.kaggle.com/datasets/meirnizri/covid19-dataset

https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico

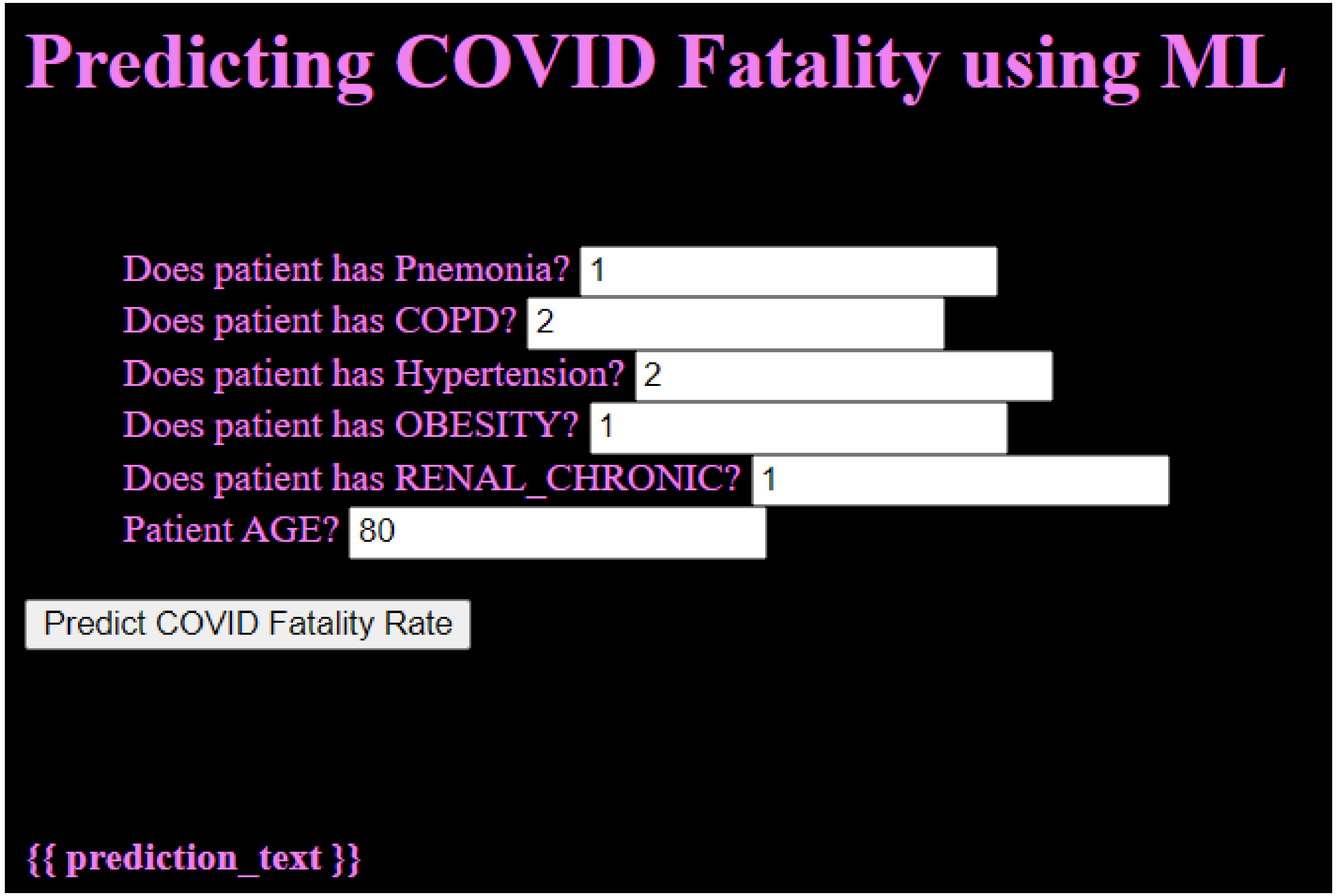
Software/Tool used:

- ✓ HTML
- ✓ Tableau
- ✓ Python
- ✓ Pandas
- ✓ Matplotlib
- ✓ NumPy
- ✓ SciPy
- ✓ PostgreSQL
- ✓ Jupyter Notebook
- ✓ Visual Code Studio

Website Playground

Tableau visualization

https://rvroomiii.github.io/group_hub/



Predicting COVID Fatality using ML

Does patient has Pnemonia? 1
Does patient has COPD? 2
Does patient has Hypertension? 2
Does patient has OBESITY? 1
Does patient has RENAL_CHRONIC? 1
Patient AGE? 80

{{ prediction_text }}