Covid19 Death Prediction

A Data Analysis and Machine Learning Project to learn about the death risk given patients' data

Duration: Jan 2023 - Mar 2023

Project Repo: https://github.com/CarynOoi/covid19_death_prediction

Python, pandas, scikit-learn, Data Extraction, Data Cleaning and Preparation, Data Visualisation

About the Project

The project provides hands-on experience in running through the process of machine learning, from data extraction, data cleaning, and data quality planning, to feeding data into the model, training, evaluating and optimising the model. The data comes from the Centers for Disease Control and Prevention (CDC: https://covid.cdc.gov/covid-data-tracker/), a health protection agency and is in charge of collecting data about the COVID-19 pandemic, and in particular, tracking cases, deaths, and trends of COVID-19 in the United States.

The three models compared in this project are linear regression, logistic regression and random forests model. The models are evaluated through cross-validation. The random forest model has the highest accuracy of 91% in this case and is further optimised by eliminating features to train the model, which reduced 95% model processing time while retaining the same model accuracy.

Some Parts of the Projects

Final features used to predict death risk for the model:

feature	importance
age_group_4	0.196655
hosp_yn_1	0.128711
age_group_2	0.071268
hosp_yn_0	0.066208
age_group_1	0.026717
case_year_2020	0.024428
hosp_yn_999	0.021631
age_group_3	0.019724
icu_yn_1	0.018925
case_year_2022	0.016726
county_fips_code_6037.0	0.015710
ethnicity_999	0.014628

Data visualisation to explore the relationship between ICU admission and death:

icu_yn
Bar Plot: