Project Report

GitHub URL

https://github.com/RadkaFahey/UCDPA_RadkaFahey

Abstract

My project is on Covid-19 mortality prediction, I looked at clinical data that could predict patients' mortality from Covid-19. I also touched on the Covid-19 pandemic data evaluation.

Introduction

I chose this project because I am interested in clinical markers for diseases. They are important for disease diagnosis and provide insight into the disease pathogenesis.

Covid-19 pandemic has hit globally in the beginning of 2020 and it has significantly impacted our lives. The main issue with infection with SARS-CoV-2, virus causing Covid-19, especially early in the pandemic, was high mortality. Being able to predict which patients need more and timely care is crucially important. Good data analysis of well-designed datasets is important in evaluating pandemic population data.

Dataset

Covid-19 mortality [1]

https://www.kaggle.com/datasets/kikemura7/covid19-patients-biomarkers?resource=download

- main dataset, published by Ikemura (2021)
- this dataset contains 4313 patient data on Covid-19 mortality and 52 columns with clinical information, including clinical laboratory test results, physical signs/symptoms and demographics data. This set is very robust regarding the amount of measured data despite some missing data.

Additional datasets to complement the main data and demonstrate some learning outcomes:

Covid-19 history [2]

https://disease.sh/

- progression of Covid-19 pandemic in time
- numbers of cases, deaths and recoveries

Covid-19 cases worldwide [3]

https://www.kaggle.com/datasets/imdevskp/corona-virus-report?resource=download

- worldwide data from July 2020 on number of confirmed, dead and recovered Covid-19 cases

Implementation Process

1. Data importing, exploration and cleaning

I imported packages needed for importing and analysis, visualization, models and machine learning.

1.1. Covid-19 mortality dataset

I manually downloaded Covid-19 mortality file. Using pandas, I read downloaded covid_dataset_original.csv file and named it covid_mortality. I looked at its shape using **.shape** function and types of the data in columns using **.dtypes** function. Death column is integer, set as binary classifier: dead=1 (yes) and death=0 (no, alive). This is useful for classification into these groups and also comparison (e.g. mean value close to 1 – more dead patients in the selected group). Age column is not integer. Authors pooled all patients older than 89 years into 89+. I replaced 89+ with 89 using function **.replace()** and made age column integer using function **.astype(int)**. To get insight into some clinical data, I sorted values using **.sort_values()** function: gender and age relating to death, and as it looks there are some differences, I subset these selected columns and named it covid_mortality_subset. I selected 10 most influential variables from Ikemura (2021) (age, systolic and diastolic blood pressure, Charlson comorbidity score, D-dimer level, pulse oximetry level, respiratory rate, blood urea nitrogen level, lactate dehydrogenase level and troponin level) and saved as covid_paperselected_ML. As patients are unique, there is not duplication, but to demonstrate the option, and to see one patient from each age, I dropped duplicates from age column using function **.drop_duplicates**. I checked null values in columns using **.isnull().sum() function**. If I remove them all I have no patients left, so I specifically removed nulls from column 'Time_from_COVID_positive_to_death_in_days' column using **.dropna()** function keeping only dead patients. For illustration, I filled in the null values with 0 using **.fillna(0)** function, but that could invalidate the results (0 value could be valid number in some columns and every patient is unique so cannot replace with any value from other patients). I checked and removed null values in datasets used for machine learning (covid_paperselected_ML).

1.2. Covid-19 history dataset

I downloaded Covid-19 history data using API and requested **.json()** which I converted into the DataFrame using pandas. I visualized the dataset using **.head()** to see first column with dates is missing heading. Using **.columns** function I have added name 'date' and named dataset covid_history. I saved it as covid_history.csv. I looked at the shape and type of the data. I added 'deaths per cases' column calculated as 'death' column/ 'cases' column to see virus mortality over the time.

1.3. Covid-19 cases worldwide dataset

I manually downloaded files on Covid-19 cases worldwide data. Using pandas, I read the downloaded country_wise_latest.csv and worldometer_data.csv and named it covid_cases1 and covid_cases2. I looked at the shapes using **.shape** function. Using **.merge** function, I joined these two datasets on common column 'Country/Region' with inner join (default) and outer join (how='outer') and named final datasets covid_cases12 and covid_cases12_outer. I compared shapes of both joined tables using **.shape** function. Using **.duplicated()** and **duplicated().sum()** functions, I looked if I have any duplicates, and there were none. I looked at list of all columns using **.columns** function.

2. Data analysis and visualization

2.1. Covid-19 mortality dataset

Using covid_mortality dataset, I did some brief stats to see the data values across columns and saved it as csv for further looking (covid_mortality_stats.csv). I looked at the mean and standard deviation of clinical data in relation to the death column using **.groupby()** function with **.agg().** I saved this dataset for further looking as covid_mortality_death.csv. I plotted 10 most influential variables from Ikemura (2021) using **seaborn sns.boxplots**. I observed median as well as spread of the values complementing my mean and standard deviation results.

Using covid_mortality_subset I looked at interrelationship of age, gender and mortality, obtaining mean values with standard deviations, using **.groupby()** function with **.agg()** and calculated number of patients in each group using **.value_counts()** function and complemented with **sns.countplot()**. I used the same functions and plots to look at effect of race and ventilator on mortality. I looked closely into relationship of age and days from the Covid-19 positive test tlll death using **scatterplot()**. I added more details into this relationship using **sns.relplot()** such as separated genders and colour for race; colored gender groups with sized ventilator use and coloured gender groups split by the number of ventilator use into each group (0-4) and sized by race.

2.2. Covid-19 history dataset

I defined function to plot time-series with specific parameters (**def plot_timeseries**). Using covid_history dataset, I used this function and **matplotlib plt.subplots()** to plot in numbers of cases and deaths per cases together, each line in different colour with a separate scale.

2.3. Covid-19 cases worldwide dataset

Using **.apply(np.sum), .apply(np.mean) and .apply(np.max)** NumPy functions I looked at summary, average and maximum values in selected columns (Tot Cases/1M pop, Deaths/1M pop and Deaths / 100 Cases) in covid_cases12_outer. I defined **lambda x function** which added column called 'Pandemic_effect' into this table with values generated by **conditional statements if else** using values from these selected columns comparing them with average values in the dataset. I have selected columns of interest from the resulting set, dropped null values using **.dropna()** function and sorted values based on these columns of interest (primarily by Deaths/1M pop) using **.sort_values()** function.

3. Data modelling / machine learning

I used covid_paperselected_ML dataset prepared earlier. I used **seaborn sns.pairplot** to see relationship of the 10 selected variables. I used **classification** machine learning model **to predict if the patient will live or not based on the selected variables**. I have defined machine learning model, decision tree classifier, generated accuracy score, confusion matrix, and applied K nearest neighbours, Linear discriminant analysis and Naïve Bayers. I have then evaluated model performance using **K-fold cross validation** and final **cv_results.mean()**.

Based on the results from statistical evaluation I selected 8 variables: age, time on ventilator, diastolic and systolic blood pressure, estimated glomerular filtration rate (egfr), ferritin, average size of platelets (mpv), and Charlson comorbidity score. I checked and dropped null values, saved as covid_selected1_MLfinal and performed same modelling and machine learning analyses as on the covid_paperselected_ML dataset.

Then I selected 4 variables, common in both selections, age, diastolic and systolic blood pressure and Charlson comorbidity score, checked for null values to be dropped and saved as covid_commonselected_ML, and performed same modelling and machine learning analyses as on the covid_paperselected_ML dataset.

Results

Covid-19 mortality data.

Figure 1 shows boxplots of 10 most influential variables on mortality from Ikemura (2021).

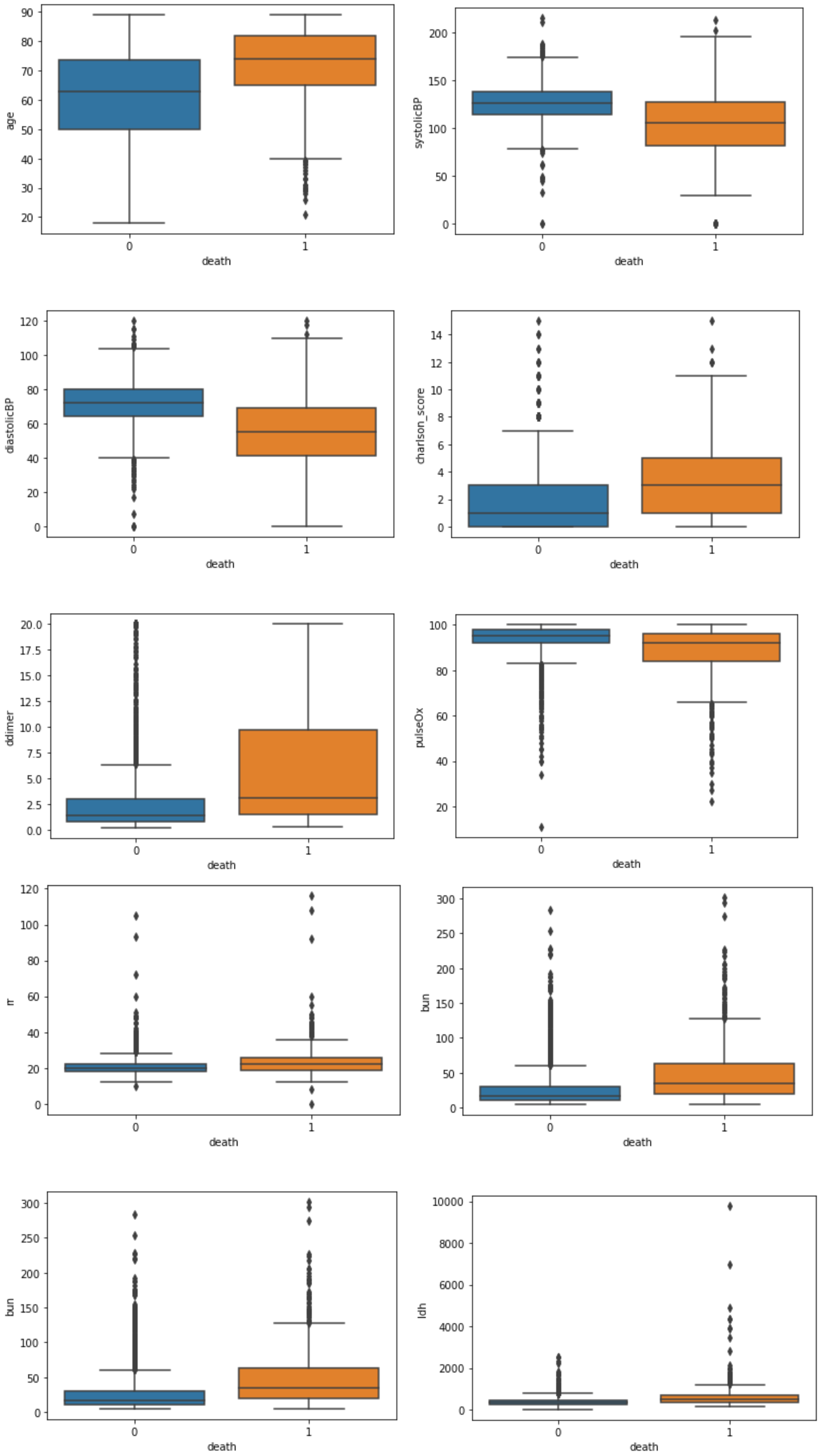
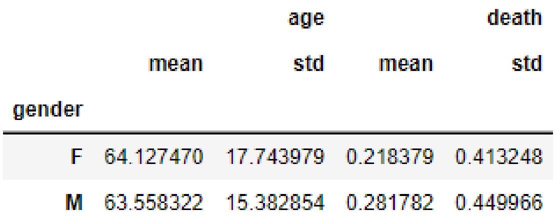


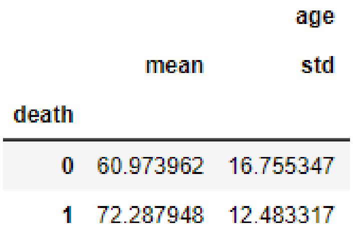
Figure 1. Boxplots of 10 variables in relation to death group (0=alive, 1=dead).

Interrelationship of age, gender and mortality.

Sorting and grouping together age and gender in relation to mortality and calculating mean values with standard deviations showed that men and older patients are more likely to die from Covid-19. (**Output [55] and [41]**):



	age		death	
	mean	std	mean	std
gender				
F	64.127470	17.743979	0.218379	0.413248
M	63.558322	15.382854	0.281782	0.449966



	age	
	mean	std
death		
0	60.973962	16.755347
1	72.287948	12.483317

There were more men in both groups (**Figure 2**).

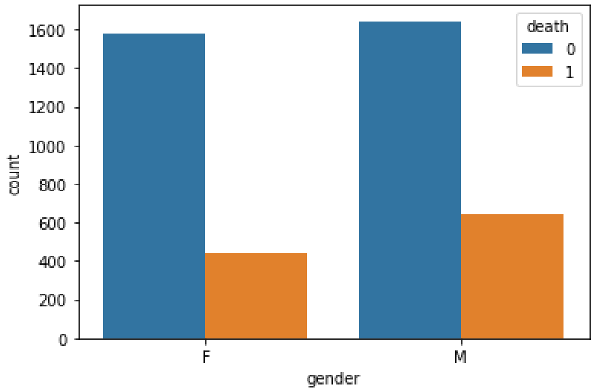


Figure 2. Countplot of gender distribution in death groups.

The age effect on how quickly the patients died since the positive test was also quite marked on the scatterplot (**Figure 3**).

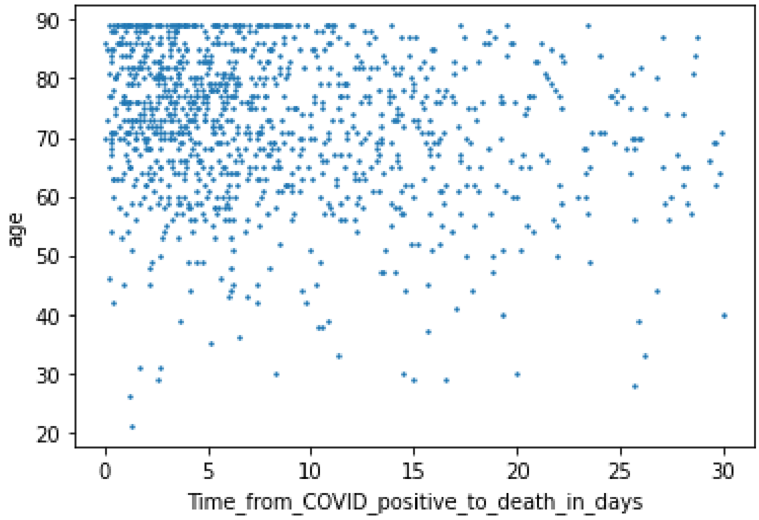
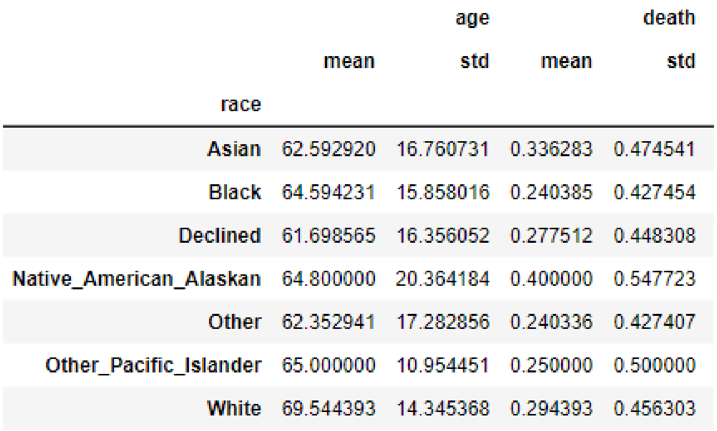


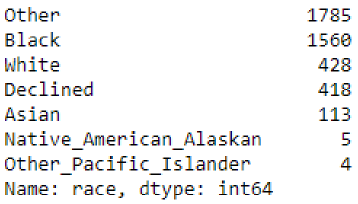
Figure 3. Scatterplot of age vs time from covid positive test to death in days.

Race and mortality

Race didn't affect mortality in this dataset, only Native American and Alaskan people have higher death rates, but there were only 5 patients in this group - too small to draw any conclusion. (**Output [58] and [59]**):



	age		death	
	mean	std	mean	std
race				
Asian	62.592920	16.760731	0.336283	0.474541
Black	64.594231	15.858016	0.240385	0.427454
Declined	61.698565	16.356052	0.277512	0.448308
Native_American_Alaskan	64.800000	20.364184	0.400000	0.547723
Other	62.352941	17.282856	0.240336	0.427407
Other_Pacific_Islander	65.000000	10.954451	0.250000	0.500000
White	69.544393	14.345368	0.294393	0.456303



```
Other                       1785
Black                       1560
White                        428
Declined                     418
Asian                        113
Native_American_Alaskan        5
Other_Pacific_Islander         4
Name: race, dtype: int64
```

This is in agreement with the scatterplot, where is no clear clustering of race (**Figure 4**).

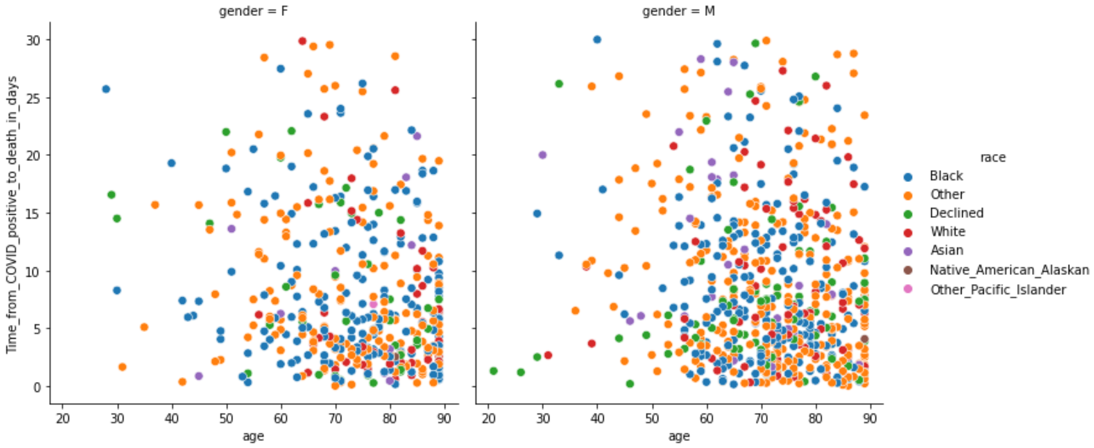
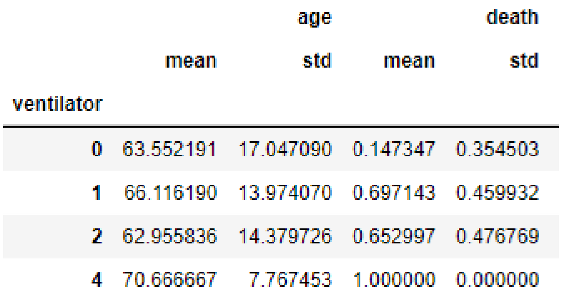


Figure 4. Scatterplot of age vs time from covid positive test to death in days, separated by gender, coloured by race.

Ventilator use and mortality

Ventilator was used more among patients which didn't survive. (**Output [60], Figure 5**)



	age		death	
	mean	std	mean	std
ventilator				
0	63.552191	17.047090	0.147347	0.354503
1	66.116190	13.974070	0.697143	0.459932
2	62.955836	14.379726	0.652997	0.476769
4	70.666667	7.767453	1.000000	0.000000

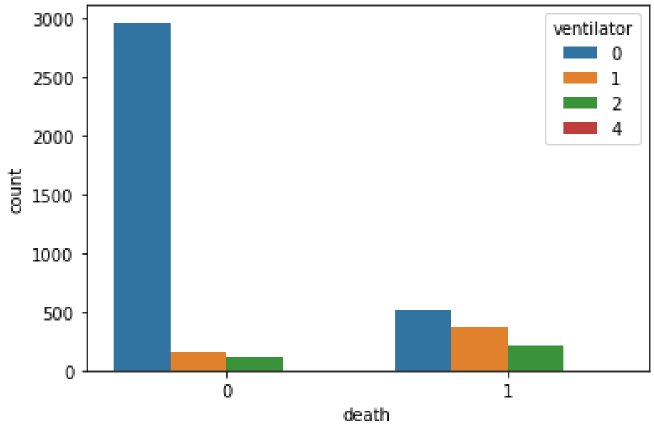


Figure 5. Countplot of dead and alive patients in relation to ventilator use.

Separating the plots based on ventilator use, it looks that ventilator did help to prolong lives of these patients (**Figure 6**).

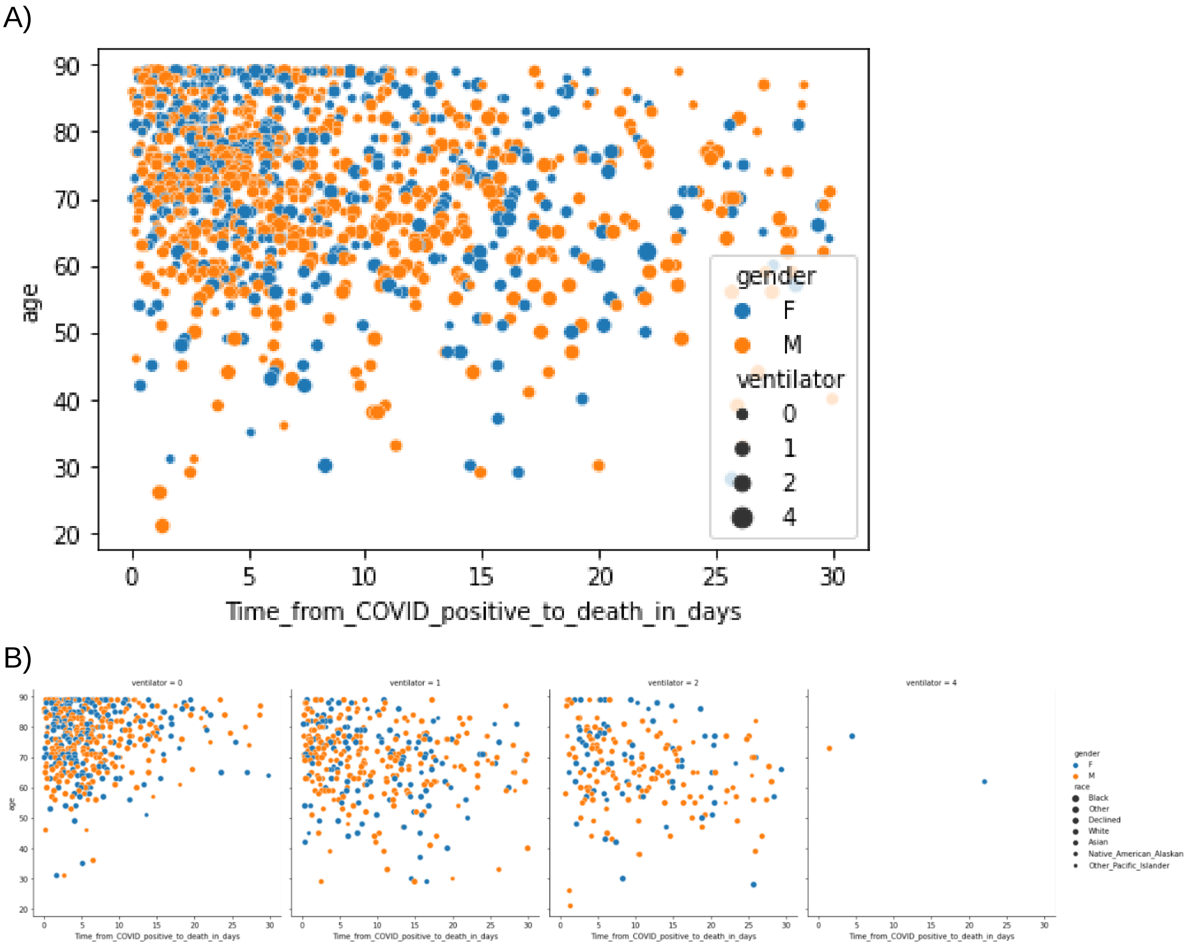


Figure 6. Scatterplot of age vs time from covid positive test to death in days, A) added gender and ventilator use and B) split by ventilator use, added race as dot size.

Machine learning to predict mortality

Machine learning classification method was used to predict mortality.

Model 1 consisted of 10 most influential variables from Ikemura (2021) and had mean predictive value of 0.86.

Model 2 consisted of my 8 selected variables (age, ventilator, diastolicBP, systolicBP" egfr,

ferritin, mpv and charlson_score) and had mean predictive value of 0.87.
Model 3. 4 common variables from model 1 and 2 were selected into model 3 and it had mean predictive value of 0.83.

Figure 7 shows sns pairplot with 4 selected variables.

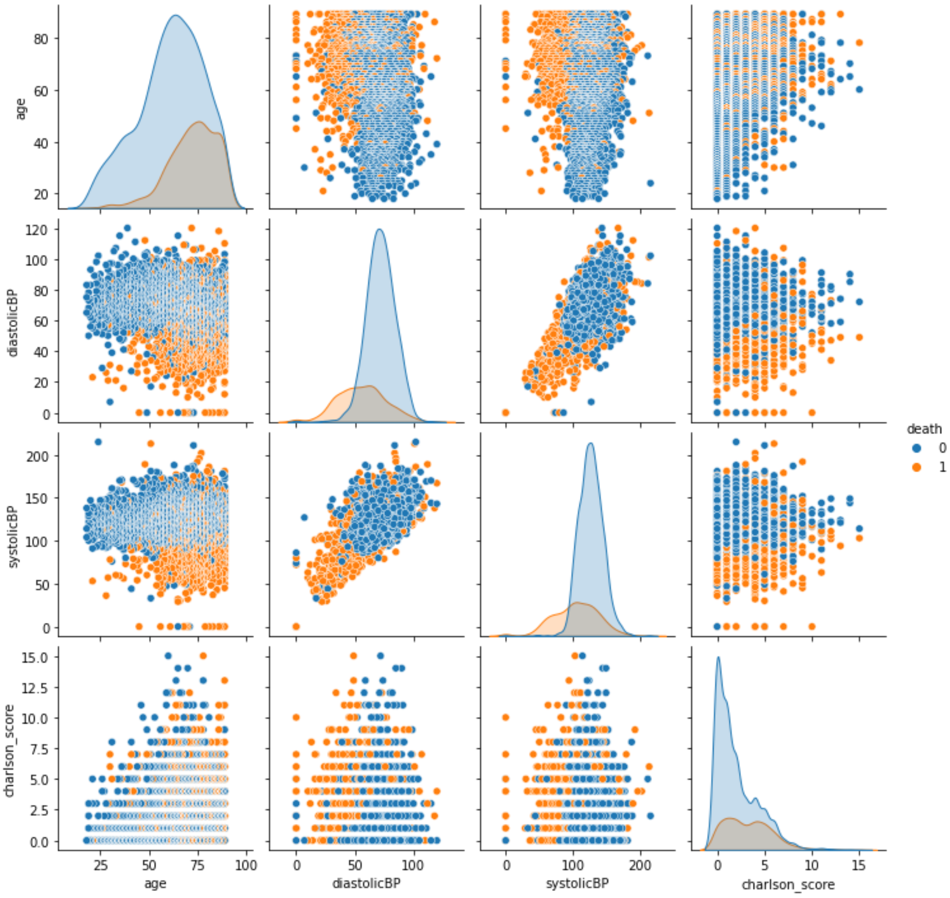


Figure 7. Pairplot of 4 selected variables in model 3 relating to death group.

Covid-19 history data.
Plotting death per cases ratio indicated that the mortality of Covid-19 goes down, while cases are still going up (**Figure 8**).

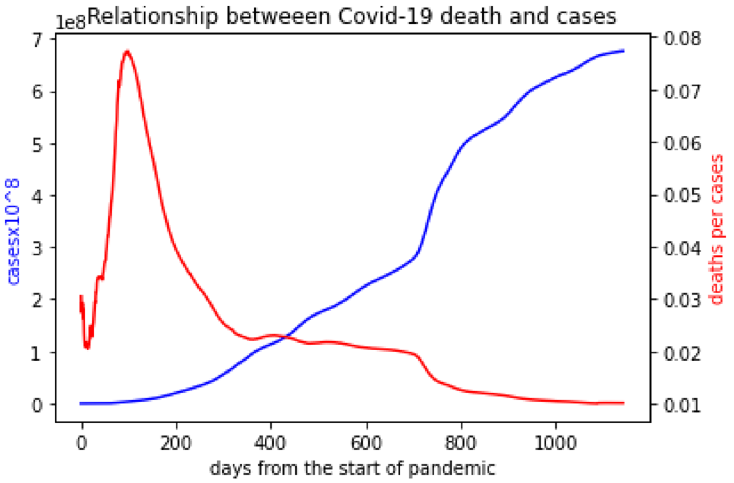
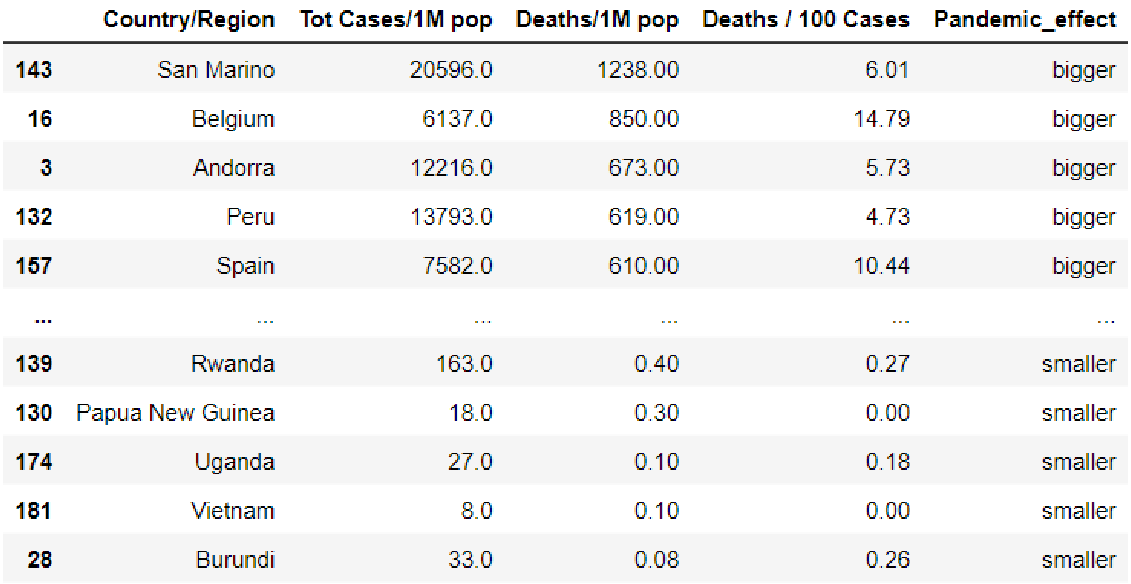


Figure 8. Line plot of Covid-19 cases in time and associated mortality.

Covid-19 worldwide cases data.

In July 2020, the top 5 countries with highest death per 1M of population and with bigger of Covid-19 pandemic effect than average were San Marino, Belgium, Andorra, Peru and Spain. Countries with lowest values of death per 1M of population and smaller than average pandemic effect were Rwanda, Papua New Guinea, Uganda, Vietnam and Burundi. (**Output [77]**):



	Country/Region	Tot Cases/1M pop	Deaths/1M pop	Deaths / 100 Cases	Pandemic_effect
143	San Marino	20596.0	1238.00	6.01	bigger
16	Belgium	6137.0	850.00	14.79	bigger
3	Andorra	12216.0	673.00	5.73	bigger
132	Peru	13793.0	619.00	4.73	bigger
157	Spain	7582.0	610.00	10.44	bigger
...	...	...	...	...	...
139	Rwanda	163.0	0.40	0.27	smaller
130	Papua New Guinea	18.0	0.30	0.00	smaller
174	Uganda	27.0	0.10	0.18	smaller
181	Vietnam	8.0	0.10	0.00	smaller
28	Burundi	33.0	0.08	0.26	smaller

Insights

- Covid-19 mortality is slightly higher in men than women (Covid-19 mortality dataset, **Output[55])**

- Age is important factor in prediction of Covid-19 mortality (**Output[41]**, it also correlates with time from the diagnosis till death (**Figure 3**)

- Race has no impact on Covid-19 mortality in this dataset (**Output [58] and [59]**) and **Figure 4**)

- Ventilator was used more among patients which died (**Output [60], Figure 5**), but ventilator did help to prolong lives of these patients (**Figure 6**)

- The best prediction biomarkers are age, time on ventilator, diastolic and systolic blood pressure, egfr, ferritin, mpv and Charlson score (0.87). However, if we only select four of these variables (age, diastolic and systolic blood pressure and Charlson score), we get only slightly reduced model performance (0.83) (**Machine learning model 1** and **3**)

- Severity of Covid-19 has decreased over the time (**Figure 8**)

- San Marino, Belgium, Andorra, Peru and Spain were mostly affected by the pandemic and Rwanda, Papua New Guinea, Uganda, Vietnam and Burundi are least affected by Covid-19 pandemic based on deaths per 1 million of population in July 2020 (Covid-19 worldwide cases joined dataset, **Output [77]**)

References

Research papers.

Ikemura, K., Bellin, E., Yagi, Y., Billett, H., Saada, M., Simone, K., Stahl, L., Szymanski, J., Goldstein, D.Y., Gil, M.R. (2021) ‘Using Automated Machine Learning to Predict the Mortality of Patients With COVID-19: Prediction Model Development Study’ *JOURNAL OF MEDICAL INTERNET RESEARCH.* 23(2), e23458.

Dataset references.

[1] Covid-19 mortality data: https://www.kaggle.com/datasets/kikemura7/covid19-patients-biomarkers?resource=download (Kaggle)

[2] Covid-19 history data: https://disease.sh/

[3] Covid-19 worldwide data: https://www.kaggle.com/datasets/imdevskp/corona-virus-report?resource=download