OCMP5310 Project Stage 1 Report

Title: Predicting ventilator requirements for COVID-19

Author: *Timothy Creer*

Date: *15th March 2025*

Section 1: Problem

1.1 Overview:

COVID-19, as defined by the World Health Organisation (WHO), is a 'symptomatic disease caused by SARS-CoV-2' (World Health Organisation, 2025). COVID-19 most commonly causes fever, chills, & a sore throat in infected individuals. As of 23/02/2025, COVID-19 has killed 7,090,776 people (World Health Organisation, 2025). As COVID-19 was a highly infectious disease, the infected population quickly overwhelmed countries' medical systems, resulting in shortages of medical equipment, particularly ventilators (Fauci, Lane & Redfield, 2020). When a person with a compromised immune system or when the body's immune system fails to fight off COVID-19, it can spread to the lungs, causing respiratory issues, making it hard for the person to breathe & possible death. To alleviate this issue, a ventilator is required to keep the person alive so the body can fight off the disease (MacMillan, 2020). As the WHO states, "People who have pre-existing health problems are at higher risk when they have COVID-19… These include people taking immunosuppressive medication; those with chronic heart, lung, liver, or rheumatological problems; those with HIV, diabetes, cancer, obesity, or dementia." If ventilator requirements could've been accurately predicted & supplied this could have saved many lives, reduced strain on the medical system & assisted policy makers during the pandemic.

1.2 Research Question:

Can we predict ventilator requirements for COVID-19 patients based on their medical history & initial symptoms?

1.3 Stakeholders:

- **Healthcare providers (Doctors, Nurses, & Hospital Administrators)**: Hospitals systems were overwhelmed during COVID-19 surges, predicting ventilator requirements allows optimal allocation of critical resources ensuring patients who need it receive them, & enables hospitals to undertake proactive decision making.
- **Public Health Officials & Government Agencies**: Having accurate predictions allow public health departments to prepare for COVID-19 spikes by having adequate ventilator stockpiles in areas where they are needed most. It also allows policy makes to make more optimal healthcare policies through better allocation of time, money & resources
- **Patient & high risk individuals**: High-risk patients (diabetes, heart disease, or immunosuppressive conditions) can benefit from early intervention strategies, reducing the likelihood of severe disease progression & help guide treatment plans.
- **Medical Device Companies**: If companies producing ventilators can better predict ventilator requirements, they can then forecast their demand more accurately thus reducing shortages or preventing surpluses in ventilators in the health system.
- **Medical Researchers & Data Scientists**: The models & insights developed from COVID-19 could also help improve treatment protocols for illnesses similar to COVID-19 & future pandemics.

Section 2: Data Overview

2.1 Data Provenance:

The dataset was originally found on kaggle under the title 'COVID-19 Dataset' which states it was acquired from the Mexican government website. On the Mexican government website it states the data is from the 'Epidemiological Surveillance System for Viral Respiratory Diseases' & the data was 'subject to validation by the Ministry of Health through the General Directorate of Epidemiology. The information contained corresponds only to the data obtained from the epidemiological study of suspected cases of

viral respiratory disease at the time they are identified in the medical units of the Health Sector'(Gobierno de México, 2025). The dataset as per the website states it was last modified on 06/08/2020.

2.2 Data License:
The dataset is governed by the "Términos de Libre Uso MX" license, which permits users to: Make & distribute copies of the dataset & its content. Disseminate & publish the dataset.
This license ensures that the data can be freely used, shared, & redistributed, promoting transparency & accessibility.

2.3 Data Structure & Metadata:

- Instances = 1,048,576
- Attributes (columns / features) = 21

Section 3: Data Quality & Cleaning

3.1 Data Ingestion: File was in a CSV format, Pandas (data handling), NumPy (missing values), Matplotlib & Seaborn (visualisation). Developed in Google Colab.

3.2 Data Quality Checks & Cleaning: Several data quality issues were found during the initial exploration & resolved as follows:

1. **Column Renaming**: Standardised columns names & fixed spelling for clarity
2. **Converting NaN**: The dataset documentation indicated that 97 & 99 represented missing values. However, 98 was also found to be missing & was converted to NaN.
3. **Dropping Irrelevant Column**s: Removed MEDICAL_UNIT & USMER as significance is unknown, & HOSPITALISED as it was unrelated to past medical conditions. DATE_DIED was also removed as it is an outcome of covid not predictor.
4. **Correcting data formats**: Converted 2 (No) to 0 in medical condition columns & encoded SEX as Male = 1, Female = 0 then set all from float64 to int64 to improve readability & compatibility with machine learning models.
5. **Correcting Pregnancy Data**: Set PREGNANT = 0 for male patients, NaN was not appropriate as the correct value should be "No" rather than "unknown."
6. **Filtering out Unrelated Data**: Excluded non-COVID cases (COVID score > 3) as per document description, greater than 3 indicates the patient is not a carrier of covid or that the test was inconclusive. Removed as we only want data related to patients confirmed with COVID.
7. **Handling Missing Data**:
    a. For INTUBATED (predictor variable) missing values were first imputed based on DATE_DIED, assuming that if a patient died from COVID-19, they likely required a ventilator before. While this is a general assumption, it helped reduce missing values from 71.99% to 70.40% but could potentially lead to false positive data.
    b. Remaining missing values for INTUBATED (70.40%) & other missing data points were dropped, as the final dataset still contained 108,160 records, which is sufficient for model development.
8. **Duplicate values**: dataset could potentially contain duplicate values however because there is no unique patient ID there is no way to check this.
9. **Outliers**: Initially the age category seemed to have some outliers due to possible entry errors as the max age was 121 which is impossible. However after initial data cleaning the age max fell to 105 is more likely but to be investigated further in EDA.

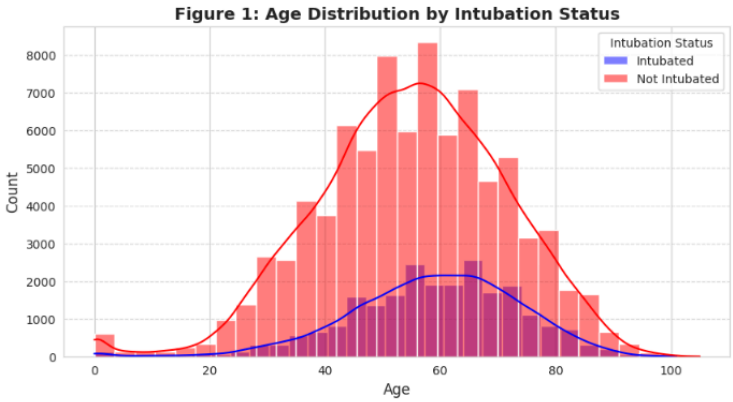
3.3 Tools Used:

- **Pandas:** pd.read_csv(), .rename(), drop(), .replace(), loc[], isna(), nunique(), dropna(), isnull(), sum(), merge(), select_dtypes(), astype(), to_frame()
- **NumPy:** np.nan (for handling missing values)

Section 4: Exploratory Data Analysis (EDA)

4.1 Exploratory Insights: EDA was conducted to further investigate outliers & identify key patterns & relationships relevant to predicting ventilator requirements for COVID-19 patients

Outlier Exploration: The AGE column, with a maximum value of 105 (as identified using .describe()), was further examined using a histogram (see Figure 1), segmented by intubated status. The histogram revealed a distinct spike at age 0, which aligns with COVID-19's higher impact on infants due to their underdeveloped immune systems. However, since newborns typically have fewer pre-existing conditions, their inclusion could introduce model bias as it is likely their intubation status could be due to other neonatal factors rather than covid. The spike at age 0 in the distribution suggests an overrepresentation that could distort feature-target relationships, weakening the predictive power of comorbidities & misclassifying adult risk levels. To improve model accuracy in later stages it was decided to remove the ages <1 year old.



For older patients, a focused histogram of those aged >94 was analysed (figure 2), showing no abnormalities. Additionally, a boxplot of age by intubated status (figure 4) highlighted a higher median age for intubated patients, indicating older individuals are at greater risk & outliers were present at <20 & >100, highlighting rarer cases in younger & very elderly patients. To quantify the outlier significance, an IQR-based outlier detection identified 663 cases (0.63% of the dataset) as outliers. Given that these outliers represent a small fraction of the dataset & also reflect realistic age distributions for COVID-19 mortality, they were retained in the analysis.

Trends:
To investigate the relationship between intubated patients & comorbidities (including ICU) a bar chart was used to visualise this relationship (figure 5). Pneumonia is the most significant risk factor, followed by hypertension, diabetes, obesity, & renal chronic disease. However conditions such as pregnancy, COPD, & asthma show lower frequencies, potentially due to dataset characteristics or treatment interventions. From this graph we can see the key features such as pneumonia, hypertension, diabetes, & obesity could significantly contribute to intubation, making them key predictive features in the future machine learning model.

To assess the overall distribution of health risks in the population, pie charts were used for each feature (figure 6) Key insights include: intubation was required for 21.52% of patients, indicating it is uncommon but still significant. Comorbidities such as pneumonia (66.58%), diabetes (31.43%), hypertension (35.3%), & obesity (23.17%) were prevalent. Additionally, the COVID severity rating was highly skewed toward

category 3 (96.57%), which may limit its predictive value. These pie charts align with findings of key features significance going forward with feature selection & model building.

A correlation analysis of all columns to Intubation was carried out revealing ICU admission (0.29), pneumonia (0.14), & age (0.11) have the strongest associations with intubation. Comorbidities such as hypertension (0.057), diabetes (0.041), & obesity (0.026) show weaker correlations. These findings highlight the importance of ICU admission, pneumonia, & age as primary predictive features while raising a key challenge of how to handle low-correlation features effectively. These variables may introduce noise, reducing model accuracy & increasing computational complexity without adding significant predictive value.

Lastly, a stacked bar chart (figure 8) was used to analyse the relationship between age & the prevalence of health issues, which confirmed that older age groups (41-60, 61-80, 81+) have a higher prevalence of pneumonia, diabetes, hypertension, & obesity. This reinforces age as a key predictive factor for ventilator requirements. The negatively skewed distribution of conditions among older adults may require age-stratified modeling rather than a single model applied to all age groups to account.

Section 5: Discussion & Conclusion

5.1 Discussion:

The dataset provided valuable insights into the prevalence of health conditions with their intubation requirements, with its large sample size ensuring statistical reliability insights. One of the main strengths is the inclusion of key medical conditions such as pneumonia, hypertension, & diabetes that are well-represented post data cleaning, allowing for meaningful future feature selection. However there are limitations, particularly with missing values (70.40% for intubation) & the highly skewed COVID severity rating (96.57% in category 3), which may limit its predictive value. The lack of unique patient identifiers means duplicate entries cannot be identified, potentially distorting patterns. Additionally, as the dataset is sourced solely from Mexico, findings may not generalise well to other healthcare systems with different medical infrastructures & patient demographics.

EDA was effective in identifying keys patterns through histograms, bar charts, pie charts, correlation analysis, & a stacked bar chart. These confirmed that there is a higher prevalence of pneumonia, diabetes, hypertension, & obesity particularly among the older ages which demonstrated a negative skew for all morbidities when grouped by ages. EDA was effective in identifying outliers in the age category along with comorbidities which have weaker predictive power like asthma, immunosuppression & renal chronic disease. Next steps will include feature engineering to account for the age skewness, low correlative features to reduce model noise & possible feature aggregation to group together mid score correlative features.

Conclusion:

The analysis identified ICU admission, pneumonia, & age as the strongest predictors of ventilator requirements. While comorbidities such as hypertension & diabetes are prevalent, their weaker correlations suggest they may introduce noise rather than add predictive value. The stacked bar chart further supports the significance of age, highlighting the need for an age-stratified modeling approach. Addressing missing data & refining feature selection will be essential for improving predictive performance in the next stages of model development.

Appendix

References

- **World Health Organization (WHO), 2025.** *WHO Coronavirus (COVID-19) Dashboard*. Available at: https://data.who.int/dashboards/covid19/cases [Accessed 12 March 2025].
- **Fauci, A.S., Lane, H.C. & Redfield, R.R., 2020.** *Covid-19 — Navigating the Uncharted. New England Journal of Medicine*, 382(13), pp.1268-1269. Available at: https://www.nejm.org/doi/full/10.1056/NEJMp2006141 [Accessed 12 March 2025].
- **MacMillan, C., 2020**. *Ventilators & COVID-19: What You Need to Know*. Yale Medicine. Available at: https://www.yalemedicine.org/news/ventilators-covid-19 [Accessed 12 March 2025].
- **Gobierno de México, 2025.** *Información referente a casos COVID-19 en México.* Datos Abiertos del Gobierno de México. Available at: https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico/resource/730779a6-6219-4839-833b-89597f524e3c?inner_span=True [Accessed 12 March 2025].

Data Dictionary:

Attribute	Description	Data Type
USMER	Indicates whether the patient was treated in medical units of the first, second, or third level.	Integer
MEDICAL_UNIT	Type of institution in the National Health System that provided care.	Integer
SEX	Patient's gender (1 = Female, 2 = Male).	Integer
AGE	Age of the patient in years.	Integer
PATIENT_TYPE	Type of care the patient received (1 = Returned home, 2 = Hospitalisation).	Integer
PNEUMONIA	Whether the patient had air sac inflammation (1 = Yes, 2 = No).	Integer
PREGNANT	Whether the patient was pregnant (1 = Yes, 2 = No, 97 = Missing Data).	Integer
DIABETES	Whether the patient had diabetes (1 = Yes, 2 = No).	Integer
COPD	Whether the patient had Chronic Obstructive Pulmonary Disease (1 = Yes, 2 = No).	Integer
ASTHMA	Whether the patient had asthma (1 = Yes, 2 = No).	Integer
INMSUPR	Whether the patient was immunosuppressed (1	Integer

	= Yes, 2 = No).	
HYPERTENSION	Whether the patient had hypertension (1 = Yes, 2 = No).	Integer
CARDIOVASCULAR	Whether the patient had a heart or blood vessel-related disease (1 = Yes, 2 = No).	Integer
RENAL_CHRONIC	Whether the patient had chronic renal disease (1 = Yes, 2 = No).	Integer
OTHER_DISEASE	Whether the patient had any other disease (1 = Yes, 2 = No).	Integer
OBESITY	Whether the patient was obese (1 = Yes, 2 = No).	Integer
TOBACCO	Whether the patient was a tobacco user (1 = Yes, 2 = No).	Integer
INTUBED	Whether the patient was connected to a ventilator (1 = Yes, 2 = No, 97 = Missing Data).	Integer
ICU	Whether the patient was admitted to an Intensive Care Unit (1 = Yes, 2 = No).	Integer
DATE_DIED	If the patient died, this field contains the date of death (Format: DD/MM/YYYY, 9999-99-99 if alive).	String
CLASIFICATION_FINAL	COVID-19 test classification (1-3 = Confirmed COVID cases, 4+ = Not a carrier or inconclusive test).	Integer