Final Project Topic Proposal

Pham Anh Khoi – ITCSIU23018
Pham Hoang Phuong – ITCSIU23056
Dang Ngoc Thai Son – ITCSIU23033

October 13, 2025

Contents

1 Problem Statement

Our group is planning to develop our final project based on some datasets about the COVID-19 pandemic. The goal is to measure the extent of the disease's impact on the patient's health, based on underlying conditions and accompanying symptoms, which may lead to fatal outcomes. Since our topic needs datasets that are not included in the suggested datasets, we have to propose our chosen datasets. The descriptions of the datasets will be provided below.

2 About Datasets

2.1 Dataset 1: COVID-19 Dataset

Link: `https://www.kaggle.com/datasets/meirnizri/covid19-dataset`
This dataset is provided by the Mexican government, contains a large amount of anonymized patient information, including details about pre-existing conditions. It consists of 21 unique features and data from 1,048,576 individual patients. A value of 1 represents "yes," and 2 represents "no," while 97 and 99 denote missing data.

- `SEX`: 1 for female and 2 for male.
- `AGE`: age of the patient.
- `CLASSIFICATION`: covid test findings. Values 1–3 mean that the patient was diagnosed with COVID in different degrees. 4 or higher means that the patient is not a carrier of COVID or that the test is inconclusive.
- `PATIENT TYPE`: type of care the patient received in the unit. 1 for returned home and 2 for hospitalization.
- `PNEUMONIA`: whether the patient already has air sac inflammation or not.
- `PREGNANCY`: whether the patient is pregnant or not.
- `DIABETES`: whether the patient has diabetes or not.
- `COPD`: indicates whether the patient has Chronic Obstructive Pulmonary Disease or not.
- `ASTHMA`: whether the patient has asthma or not.
- `INMSUPR`: whether the patient is immunosuppressed or not.
- `HYPERTENSION`: whether the patient has hypertension or not.
- `CARDIOVASCULAR`: whether the patient has heart or blood vessel-related disease.
- `RENAL CHRONIC`: whether the patient has chronic renal disease or not.
- `OTHER DISEASE`: whether the patient has another disease or not.
- `OBESITY`: whether the patient is obese or not.
- `TOBACCO`: whether the patient is a tobacco user.
- `USMR`: indicates whether the patient was treated in medical units of the first, second, or third level.
- `MEDICAL UNIT`: type of institution of the National Health System that provided the care.
- `INTUBED`: whether the patient was connected to the ventilator.
- `ICU`: indicates whether the patient had been admitted to an Intensive Care Unit.
- `DATE DIED`: if the patient died, indicates the date of death, and 9999-99-99 otherwise.

2.2 Dataset 2: Comorbidities and Symptoms of COVID-19 Patients

Link: `https://www.kaggle.com/datasets/martuza/comorbidities-and-symptoms-of-covid19-patients?`
This dataset contains 2 CSV files: `comorbidity.csv` and `symptoms.csv`. It includes information on comorbidities and symptoms of 1,143 COVID-19-positive patients, along with their demographic details and outcomes (alive or deceased).

`comorbidity.csv`

This file contains comorbidity information for each patient. The attributes are as follows:

- `sex`: Gender of the patient (1 = male, 2 = female)
- `age`: Age of the patient (in years)
- `hypertension`: Presence of hypertension (1 = yes, 0 = no)
- `cardiovascular`: Presence of cardiovascular disease (1 = yes, 0 = no)
- `cerebrovascular`: Presence of cerebrovascular disease (1 = yes, 0 = no)
- `lung`: Presence of lung disease (1 = yes, 0 = no)
- `malignancy`: Presence of malignancy (1 = yes, 0 = no)
- `diabetes`: Presence of diabetes (1 = yes, 0 = no)
- `liver`: Presence of liver disease (1 = yes, 0 = no)
- `kidney`: Presence of kidney disease (1 = yes, 0 = no)
- `neurodegenerative`: Presence of neurodegenerative disease (1 = yes, 0 = no)
- `infectious`: Presence of infectious disease (1 = yes, 0 = no)
- `surgical`: History of surgical condition (1 = yes, 0 = no)
- `copd`: Presence of Chronic Obstructive Pulmonary Disease (1 = yes, 0 = no)
- `asthma`: Presence of asthma (1 = yes, 0 = no)
- `outcomes`: Patient outcome (1 = deceased, 0 = alive)

`symptoms.csv`

This file records the symptoms of COVID-19 patients. The columns include:

- `headache`: Presence of headache (1 = yes, 0 = no)
- `fever`: Presence of fever (1 = yes, 0 = no)
- `cough`: Presence of cough (1 = yes, 0 = no)
- `fatigue`: Presence of fatigue (1 = yes, 0 = no)
- `nausea`: Presence of nausea (1 = yes, 0 = no)
- `diarrhea`: Presence of diarrhea (1 = yes, 0 = no)
- `myalgia`: Presence of muscle pain (1 = yes, 0 = no)
- `dyspnea`: Presence of shortness of breath (1 = yes, 0 = no)
- `pneumonia`: Presence of pneumonia (1 = yes, 0 = no)

- `ards`: Presence of Acute Respiratory Distress Syndrome (1 = yes, 0 = no)
- `septic`: Presence of septic condition (1 = yes, 0 = no)
- `age`: Age of the patient (in years)
- `gender`: Gender of the patient (1 = male, 2 = female)
- `outcomes`: Patient outcome (1 = deceased, 0 = alive)

3 How the Data will be Applied

Our group aims to integrate these two datasets based on the similar features between them and perform preprocessing to extract and select the most relevant features. The processed data will then be utilized to train a machine learning model in the future. The output will

- Give the probability that the patient can recover from COVID-19 or the predicted date of death if the patient can not recover.
- Identify the diseases, factor that make people susceptible to COVID.