CS5002 Final Project

Team:

Context:

Over the past 2 years, the Covid-19 pandemic has dramatically altered our daily lives and put a nearly insurmountable burden on the healthcare system. We chose to analyze a data set of patients with Covid-19. Our goal is to develop a machine learning algorithm to identify patients at higher risk of serious COVID-19 symptoms.

- For me, Covid-19 has altered countless aspects of my life and current

of the pandemic there was intense fear. Patients called the pharmacy nonstop seeking information. In addition, supply chain worries were a concern, the potential for medication shortages was real. A rush to stock up on medications overwhelmed the pharmacy workflow. As the pandemic progresses, community pharmacies have taken on additional responsibilities in combating and treating Covid-19. Currently, Covid booster shots are being given every day. Furthermore, Paxlovid, a Covid-19 treatment, is dispensed nearly every day. Thus, analyzing a Covid-19 data is of great interest to me.

- It has been three years since the outbreak of the Covid-19 epidemic. Although the harmfulness of Covid-19 has continued to decline with the evolution of the virus, there is still a certain degree of severe illness and death.

Based on this situation, I am quite interested in using the knowledge of Probability I learned in class, combined with the popular machine learning methods, to predict and explore the factors that cause the high risk of Covid-19. Furthermore, it can help identify susceptible groups of people, and then allocate medical resources more effectively and efficiently.

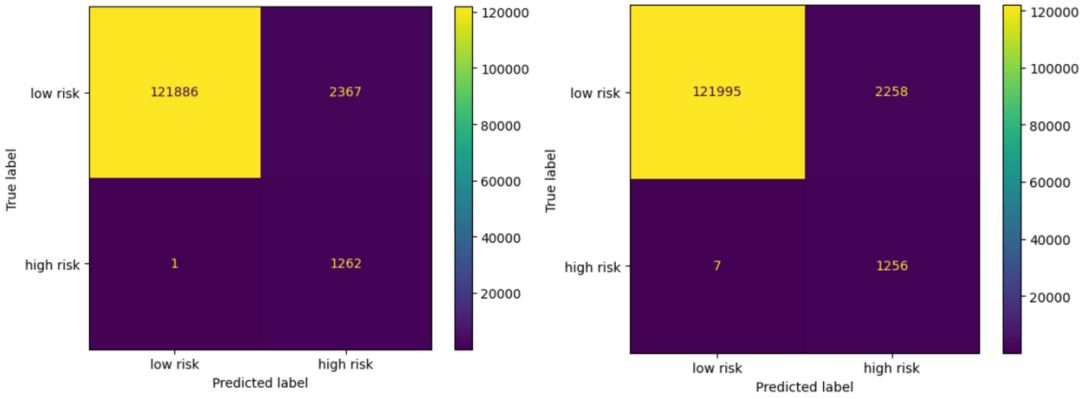
- Back in my country there were still stringent zero-Covid policies including massive lockdowns in major provinces and cities up until Nov this year. Some key concerns, for not softening the control over the past three years, are lagging vaccination rate, herd immunity yet to achieve and limited intensive care infrastructure per capita. Just a few days ago, the government announced a gradual opening up. I believe there will be outbreaks in the near future and the healthcare system could be overwhelmed if not prepared in advance. I feel interested in finding out which groups of people are more likely to get very sick after affecting the Covid virus. Are the elderly, or those with certain underlying medical conditions? Based on the data of past covid cases, we may be able to gain some insights on this so that we can protect our family, friends and ourselves better.

Application:

Our application will be a (probably ensemble) ML model that can help to predict whether a confirmed Covid patient is at high risk of death or developing severe symptoms that would require medical resources including intubation and admission to ICU.

So far we have built several potential models with Python using Naive Bayes Classifiers (including Bernoulli, Gaussian and Multinomial) and Random Forest Classifier. The models are trained with a dataset of information regarding COVID-19 cases in Mexico published by the Ministry of Health of Mexico. We pre-processed and cleaned the data first, then 80/20 split into training and test subsets, as well as ensuring the same proportions of class label as the input dataset. We applied GridSearchCV technique to look for the optimal set of parameter values which is also cross validated. To compare the test results of each model, we plotted confusion matrices and generated classification reports to get the recall and fall-out rates.

For now the prediction accuracy rates by our potential models are considerably high (close to 100%). The cause for that is probably due to the imbalance property of the original dataset. To address this problem we plan to further undersample the data and use techniques such as regularization.



Scope:

We intend to fully work through the data set. We will clean the data looking for missing values, duplicate records, and removing extraneous information. Continuing, we will do exploratory analysis with descriptive statistics and visualizations. We will look at the overall incidence of the Covid-19, potentially aggravating medical conditions (asthma, diabetes, hypertension). Next, we will develop a machine learning model to predict which patients are at high risk of severe Covid-19 complications and should be given the highest-level of care. Whereas those not likely to have severe complications can be discharged home, reducing the strain on the healthcare system. Though many other factors, like vaccination status, could impact the possibility of developing severe symptoms, we intend not to examine these factors due to the short timeframe of this project and insufficient data.

Description:

What we have done:

1. Collect data:

   Our data set was collected from the website of the Mexican government, which has been tracking information regarding COVID-19 cases in Mexico.

2. Clean data:

   The database is updated daily and contains a vast number of anonymized patient-related information since 2022. The raw dataset consists of 40 different features and 1,048,575 unique patients. Most of the features are Booleans in which “1” means "yes" and “2” means "no". The values “97” and “99” are missing data.

   Before analysis, we did some data filtering and cleaning. depending on the problem and the relationships we want to explore, we

   1. dropped those cases that were not confirmed of Covid-19;
   2. reduced the dataset to 22 columns, including “sex”, “age”, “date of death”, and medical histories of patients. Also, we filtered out those rows with missing/NA values in features.

3. Explore data:

   We have started on the analysis with descriptive statistics as well as data visualization.

4. Prepare data for ML:

   To further process data before building models, we:

   1. Transformed categorical features into binary data;
   2. Created the target variable by taking into account three attributes that we felt were dramatic contributors to severe COVID-19: whether a patient required intubation, required admission to an Intensive Care Unit or died;
   3. Split the dataset into training and testing parts with a ratio of 80% and 20%, respectively, and with the same proportions of the class label;
   4. Undersampled the training subset.

5. Build models

   We have built 3 models with Naive Bayes Classifiers (including Bernoulli, Gaussian and Multinomial) and 1 model with Random Forest Classifier.

6. Tune parameters

   We used GridSearchCV to tune and cross-validate different combinations of parameters, and generated the best estimator with optimal parameters

7. Visualize model performance

   We plotted confusion matrices and ROC curve graphs to compare prediction results

What we plan to do:

1. Over the next two weeks we plan to examine the data set further looking for correlations in our data and severe Covid-19 and create visualizations to further understand the data.
2. We plan to build new models applying other popular machine learning classification algorithms (such as Support Vector Machine, Multilayer Perceptron/Neural Network, Logistic Regression, K-Nearest Neighbors).
3. We will choose our final model, analyze those significant predictors and conclude our findings.