Group 19 Final Report

Aidan Dobratz, Robert Dreyer, Ketan Kotla, Justin Li, Dennis Luis-Aragon

Introduction

Throughout the course of the COVID-19 pandemic, one of the main problems that healthcare providers have faced is the shortage of medical resources and a proper plan to efficiently distribute them. In these tough times, being able to predict what kind of resources an individual might require at the time of being tested positive would be of immense help to medical professionals, as they would be able to more efficiently allocate resources to the patients that are most at risk of death. We set out to determine which of the features are most important for classifying if a patient has a high chance of passing away and then use these in various prediction models.

Data

The data we ended up using was Covid-19 patient data from the Mexican government found here. It included over a million rows with 21 unique features.

	USMER	MEDICAL_UNIT	SEX	PATIENT_TYPE	DATE_DIED	INTUBED	PNEUMONIA	AGE	PREGNANT	DIABETES	...	ASTHMA	INMSUPR	HIPERTENSION
0	2	1	1	1	03/05/2020	97	1	65	2	2	...	2	2	1
1	2	1	2	1	03/06/2020	97	1	72	97	2	...	2	2	1
2	2	1	2	2	09/06/2020	1	2	55	97	1	...	2	2	2
3	2	1	1	1	12/06/2020	97	2	53	2	2	...	2	2	2
4	2	1	2	1	21/06/2020	97	2	68	97	1	...	2	2	1
5	2	1	1	2	9999-99-99	2	1	40	2	2	...	2	2	2
6	2	1	1	1	9999-99-99	97	2	64	2	2	...	2	2	2
7	2	1	1	1	9999-99-99	97	1	64	2	1	...	2	1	1
8	2	1	1	2	9999-99-99	2	2	37	2	1	...	2	2	1
9	2	1	1	2	9999-99-99	2	2	25	2	2	...	2	2	2

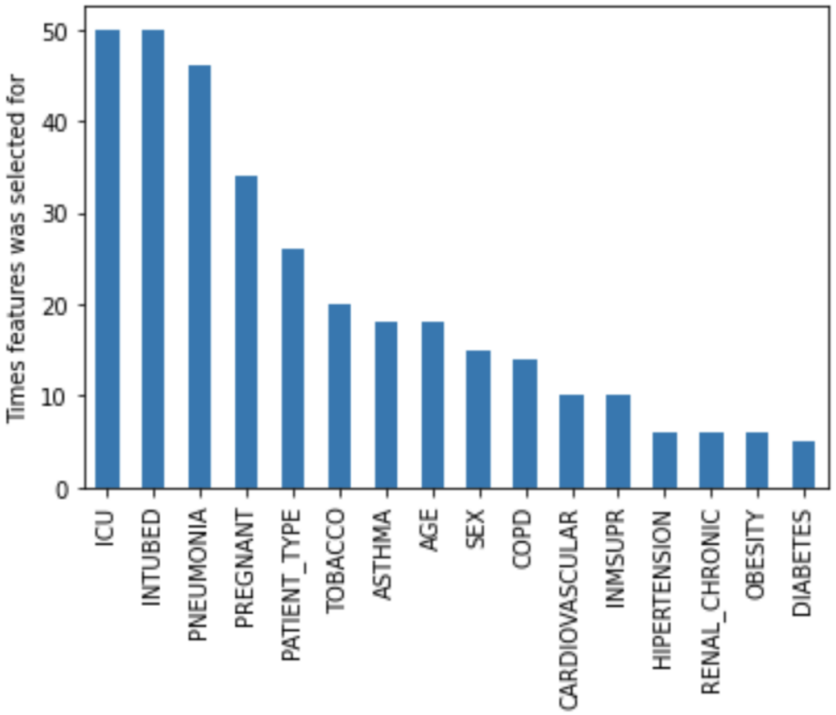
The features included information patient's pre-existing conditions such as pneumonia or asthma and how severe their illness was such as whether they were on the ventilator (intubed) or under the ICU (intensive care unit). For boolean features, 1 means yes and 2 means no and missing values are represented by 97 or 99.

Data Cleaning

Data cleaning allowed us to identify and fix errors, duplicates, and irrelevant data from the raw dataset. As mentioned before, the original dataset had over a million rows of information. With such a large number of data, there had to be missing or incorrect details. In order to generate the appropriate models and decisions, we filtered out negative or inconclusive covid test results. Also, missing data from each column was changed to -1, which is out of range from the other values in our dataset. The "Age" column needed some reworking as well, since the values were out of our [0-1] range. To solve the issue, the group applied MinMaxScalar for the Age column. This method corrected the out-of-range dilemma. As a final detail in the correction process, the "Date_Died" column was converted to "survived", in order to use it as a dependent variable.

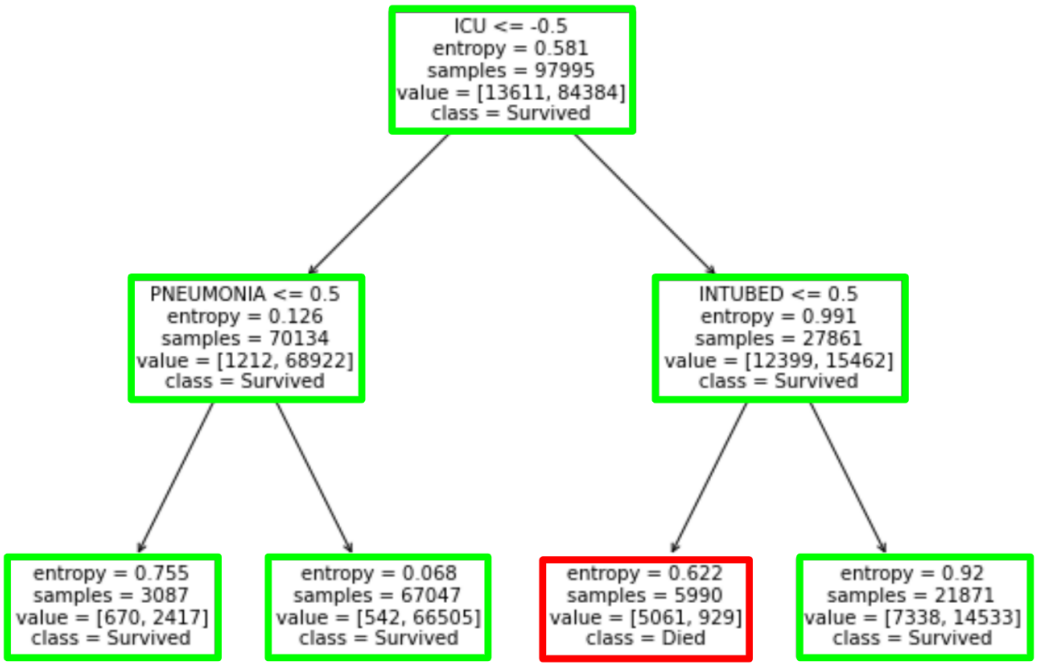
Methods

To determine which features were most predictive of COVID outcomes, we performed features selection by training a model with an L1 penalty. Initially we tried an SVM, but due to the size of our data, this took too long. Instead we used a stochastic gradient descent (SGD) classifier with L1 penalty and alpha = 0.01. Because the SGDClassifier has a random initialization, we kept getting different feature selection each time it was run. To account for this, we trained the model 50 times and counted which features were selected the most



number of times. The most selected features were: admission to the ICU, connection to a ventilator (intubed), had pneumonia, and if they were pregnant. We used these features for the models in our grid search to determine the best model.

Using the top four variables as determined by the SGDClassifier, we ran a grid search to determine the best model from logistic regression, decision tree, and KNN. For our logistic regression model we set the max iteration to 5000 and tried C values of .01, 1, and 100. Next, for the decision tree classifier we used a criterion of entropy and used max depths of 1, 3, 5, and 7. Lastly, for our K-Neighbors classifier we used a euclidean metric with n values of 1, 2, 3, and 4. The best model was a decision tree with max depth of 7 with an accuracy of .9043 on validation data. The following graph shows the decision tree at a depth of two. Green boxes represent classifications of "survived" whereas the red box represents a classification of "died". When using the top four variables, patients were most likely to die if they were admitted to the ICU and were not intubed.



Results

Using the decision tree model on our test data gave a score of approximately .91. However considering the data was 86% survivors, this isn't too impressive. After rebalancing the data with undersampling, retraining and revalidating our models showed KNN to be the best classifier. This received a score of .874 on the resampled test data. Overall this shows that these 4 features are predictive of COVID outcomes.

Contributions

Member	Proposal	Coding	Presentation	Report
Aidan Dobratz	1	.6	1	1
Robert Dreyer	1	.8	.8	1
Ketan Kotla	1	1	1	1
Justin Li	1	.7	.7	1
Dennis Luis-Aragon	1	.7	1	1