Henry McTernan & Maris Ryan

Data-Driven Insights into COVID-19 Mortality

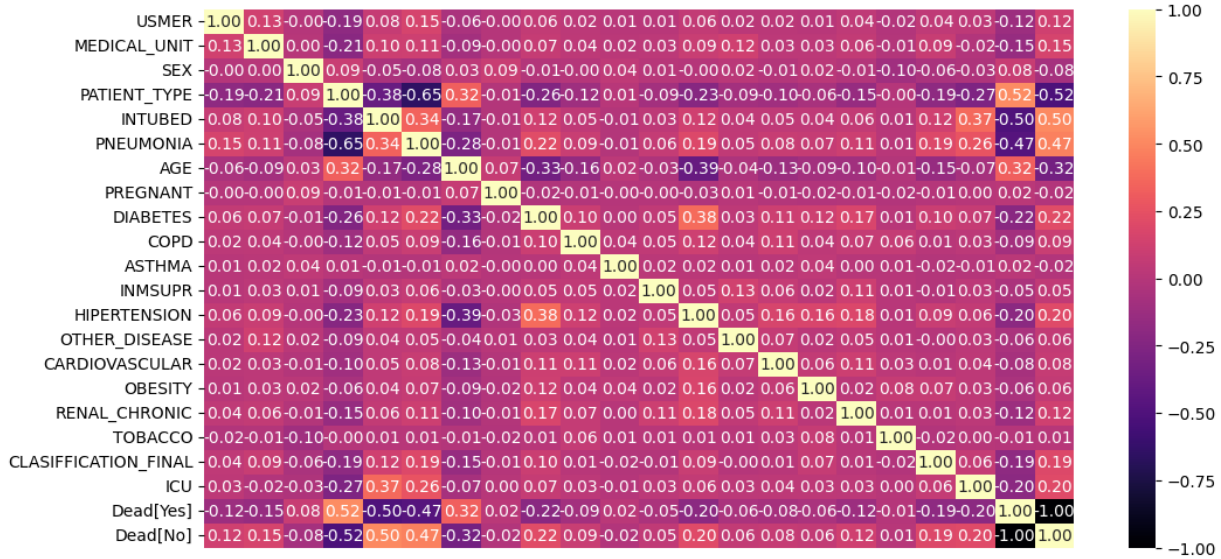
Description of research questions/issues & the significance of the problems:

COVID-19 has impacted the lives of everyone in the world since the beginning of 2020. Over a million people have died in the United States alone and almost 7 million have died worldwide. We will be researching how accurately we can predict the death of a person with COVID-19 using a trained algorithm. We will be looking into the different factors that may cause a person to be at a high risk of dying from the virus. People will be more informed about which factors lead to more worry about contracting the virus.

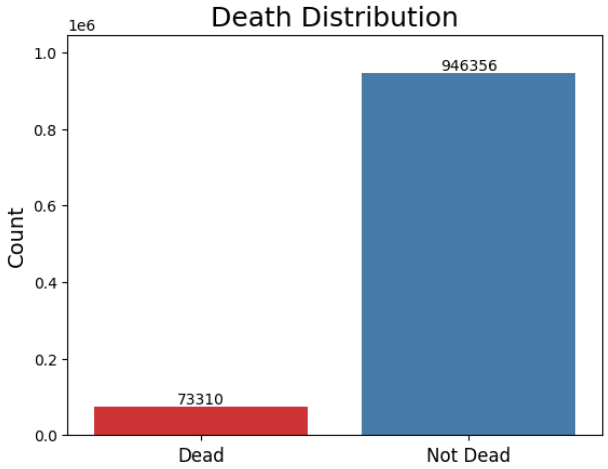
Description of data:

The COVID-19 death dataset we used contains 21 predictors. Along with the predictors, there are 1,048,576 observations (patients) provided by the Mexican government. The predictors used are USMER, Medical_Unit, Sex, Patient_Type, Date_Died, Intubed, Pneumonia, Age, Pregnant, Diabetes, COPD, Asthma, Inmsupr, Hypertension, Cardiovascular, Renal chronic, Other disease, Obesity, Tobacco, ICU, and Classification. We will use these predictors to predict if an individual will die from the CoronaVirus. USMER shows whether the patient treated medical units of the first, second, or third level. Medical_Unit expresses the type of institution of the National Health System that provided the care. Sex indicates if the patient is male or female. Patient_Type is if the patient could go home or go to the hospital. Date_Died indicates the date a patient died. Intubed shows patients' need for a ventilator. Pneumonia is whether the patient already has air sacs inflammation or not. Age is how old the patient is. Pregnant shows if the patient is pregnant. Diabetes is whether the patient has diabetes. COPD is Chronic obstructive pulmonary disease. Asthma shows if a patient has asthma. Inmsupr whether the patient is immunosuppressed. Hypertension if the patient has hypertension (when the pressure in your blood vessels is too high (140/90 mmHg or higher)). Cardiovascular whether the patient has heart or blood vessels related disease. Renal chronic whether the patient has chronic renal disease. Other diseases the patient has. Obesity indicates overweight patients. Tobacco; patients use of tobacco. ICU is the patients need to go to an Intensive Care Unit. Classification shows values 1-3 mean that the patient was diagnosed with COVID-19 in different degrees. 4 or higher means that the patient is not a carrier of COVID-19 or that the test is inconclusive. After cleaning our data, we converted the Date_Died column to classification, where class 1 means the patient died and class 0 means the patient did not die.

Preliminary studies:

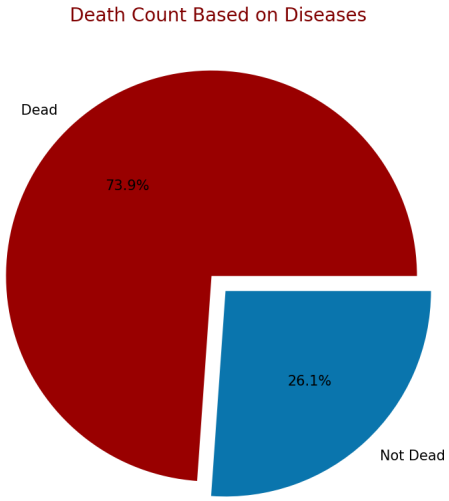


From the correlation table, we can see which variables are relevant and irrelevant for predicting the death of a patient with COVID-19. Most of our predictors are uncorrelated however some are negatively correlated like patient_type and pneumonia, hypertension and age, and dead and patient type. These make sense because patient type is a binary variable with 1 being going home and 2 being hospitalization. So, when dead is 1 then the patient is hospitalized and when dead is 0 the patient is sent home. With hypertension, it is a binary variable with 1 being having it and 2 being doesn't have it. So when a patient has hypertension they are usually older hence the negative correlation. Also, there is a high correlation between dead and intubed and dead and pneumonia. This also makes sense because if a patient is on a ventilator they have a much higher chance of death than a patient who is not on a ventilator. The same idea is for someone with pneumonia. There are not many highly correlated variables in our dataset, which indicates there is not a strong linear relationship between those specific pairs of variables.

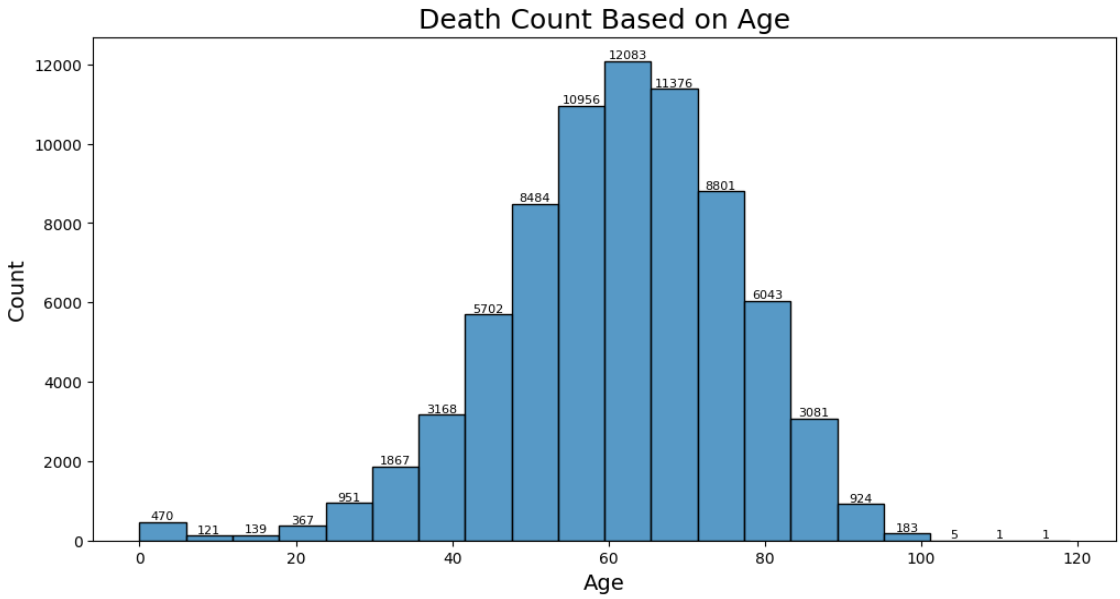


We see that roughly 7.1896% of patients in this study have died from COVID-19. While this number may seem low, when comparing it to Mexico as a whole, the death rate is only

4.3517%. So Mexico has a much lower death rate than in this dataset. This is very different from the United States where the death rate is 11.0048% from COVID-19. However, this also shows that our data is imbalanced, which we fixed later on in our analysis through resampling.



From this chart, we see that almost 74% of deaths from COVID-19 are patients that have at least one of the following diseases: pneumonia, asthma, chronic obstructive pulmonary disease, diabetes, immunosuppressed, cardiovascular disease, chronic renal disease, obesity, or the disease was labeled as "other." Indicating that as a whole, already having a disease can lead to death if a patient were to contract COVID-19.

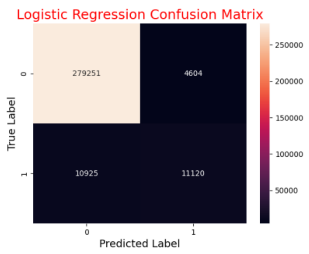


We can see that deaths from COVID-19 based on the ages of patients are almost uniformly distributed around 60 years old. This seems to make sense as the deaths from diseases

in the chart shown above are more common in older people. Since these diseases account for 74% of COVID-19 deaths, it would seem as though the older a person is the more likely that they are to contract one of these diseases and have a higher probability of dying.

Statistical analysis (methods & results): (All confusion matrices can be found in slides)

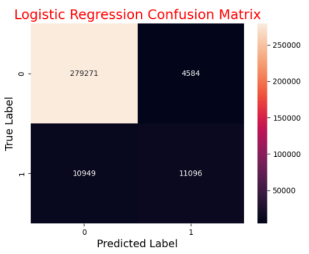
Missing values:



Accuracy = 0.9492350441320693
Precision = 0.7071991859577715
Recall = 0.5044227716035382

FPR = 0.29280081404222846
FNR = 0.03764956440229378

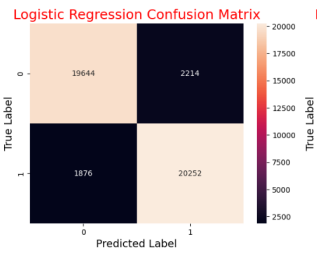
Irrelevant variables:



Accuracy = 0.9492219679633868
Precision = 0.7076530612244898
Recall = 0.5033340893626673

FPR = 0.2923469387755102
FNR = 0.037265522706912

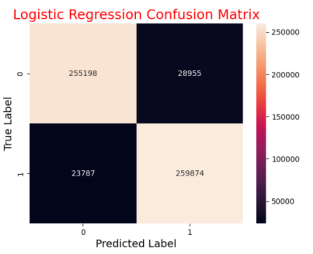
Undersampling:



Accuracy = 0.9070158686854909
Precision = 0.9014510816344699
Recall= 0.9152205350686913

FPR = 0.09854891836553013
FNR = 0.08717472118959108

Oversampling:

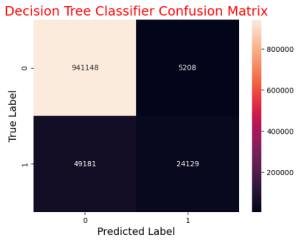


Accuracy = 0.9071139492862099
Precision = 0.899750371326979
Recall = 0.916142860668192

FPR = 0.10024962867302106
FNR = 0.08526264852949084

After missing values were dealt with and we trained the model on a split of 30% test size, we found the accuracy of logistic regression is 0.9492 with a precision of 0.7072 and recall of 0.5044. Also, the FPR is 0.2928 and FNR is 0.0376. However, after removing irrelevant variables the accuracy is 0.9492 with a precision of 0.7077 and a recall of 0.5033. The FPR is 0.2923 and FNR is 0.0377. Both of these models have practically identical results. Next, we resampled the data and viewed the results. There is undersampling, which deletes samples from our majority class (‘Not Dead’), and oversampling, which duplicates samples from our minority class (‘Dead’). When undersampling we found the accuracy to be 0.9070 with a precision of 0.9015 and recall of 0.9152. The FPR is 0.0985 and FNR is 0.0872. Then when oversampling the accuracy is 0.9071, the precision is 0.8998, and has a recall of 0.9161. The FPR is 0.1002 and FNR is 0.0853. When resampling both over and undersampling the outcomes are down across the board so these models will not be considered for our final model, but it is a good indicator of what may happen during a real-life trial.

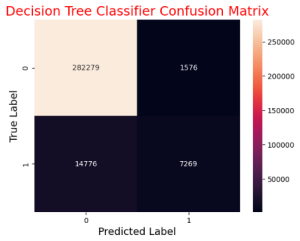
Missing values



Accuracy = 0.9466599847401012
Precision = 0.65
Recall = 0.52

FPR = 0.17752326413743735
FNR = 0.04966127418262012

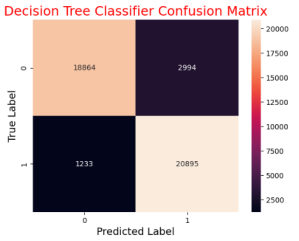
Irrelevant variables:



Accuracy = 0.9465446224256293
Precision = 0.64
Recall = 0.53

FPR = 0.17817976257772752
FNR = 0.3431889443734758

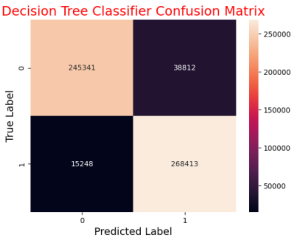
Undersampling:



Accuracy = 0.9039012413040513
Precision = 0.64
Recall = 0.53

FPR = 0.1253296496295366
FNR = 0.06135244066278549

Oversampling:

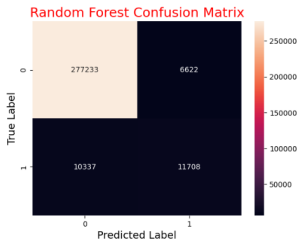


Accuracy = 0.9071139492862099
Precision = 0.899750371326979
Recall = 0.916142860668192

FPR = 0.10024962867302106
FNR = 0.08526264852949084

For the decision tree, we used the same processes, respectively, and found an accuracy of 0.9467, precision of 0.65, recall of 0.52, FPR of 0.1775, and FNR of 0.0497 (missing values). Then an accuracy of 0.9465, precision of 0.64, recall of 0.53, FPR = 0.1782, and FNR = 0.3432 (removing irrelevant variables). Next, accuracy = 0.9039, precision = 0.64, recall = 0.53, FPR = 0.1253, and FNR = 0.0614 (undersampling). Finally, accuracy is 0.9048, precision is 0.89, recall is 0.92, FPR is 0.1263, and FNR = 0.0585 (oversampling). We can see that from cleaning missing numbers and removing irrelevant variables there is again not a noticeable difference in any of the values, but when resampling all of the values decrease. Both logistic regression and decision tree have very similar results for the four processes, so it is best to try a different model.

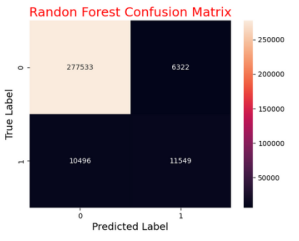
Missing values:



Accuracy = 0.9445603138280484
Precision = 0.64
Recall = 0.53

FPR = 0.36126568466993997
FNR = 0.03594603053169663

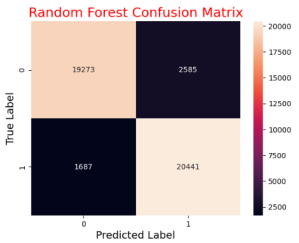
Irrelevant variables:



Accuracy = 0.9450212487741092
Precision = 0.65
Recall = 0.52

FPR = 0.3537574841922668
FNR = 0.03644077506084457

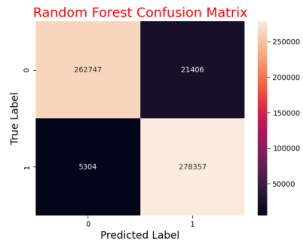
Undersampling:



Accuracy = 0.9028781885145274
Precision = 0.89
Recall = 0.92

FPR = 0.11226439676886997
FNR = 0.08048664122137404

Oversampling:



Accuracy = 0.9529599481520357
Precision = 0.93
Recall = 0.98

FPR = 0.07140974703348979
FNR = 0.019787279286404454

The final machine learning model we used was random forest classification. Again, we used the same four methods to see which gave the best results. We found accuracy to be 0.9446, precision to be 0.64, recall to be 0.53, FPR to be 0.3613, and FNR to be 0.0359. Then, we found accuracy to be 0.9450, precision to be 0.65, recall to be 0.52, FPR to be 0.3538, and FNR to be 0.0364. Next, accuracy was found to be 0.9029, precision to be 0.89, recall to be 0.92, FPR to be 0.1123, and FNR to be 0.0805. Finally, accuracy is 0.9530, precision is 0.93, recall is 0.98, FPR

is 0.0714, and FNR is 0.0198. This time the results were different, specifically with oversampling showing the best results. The accuracy was the highest, FPR and FNR were substantially lower than the other methods, and recall and precision were much higher. Overall, this would be a model that we would test on a new dataset of the same variables.

Discussion/conclusion:

COVID-19 has changed the lives of everyone around the world over the last three to four years. Predicting the risk of a patient dying from COVID-19 is a critical issue that can be solved through machine learning. The global impact that this pandemic had, and still has, implies the significance and urgency of understanding which factors influence mortality rates. We believe that our analysis can contribute valuable insight into understanding this virus better and educating individuals on the health risks. The improvement of effective prediction models can be used to apply the models to different variants of COVID-19 or other health crises. Our best model was the random forest test after oversampling. Our model would be able to accurately predict the patient's risk of dying from COVID-19 about 95.30% percent of the time. Our observed false positive rate of 7.14% could be a possible concern, however, it is better in our case of COVID-19-related deaths to have a higher false positive rate than a false negative rate. Our false negative rate is 1.98%, which is crucial in the healthcare context and emphasizes the effectiveness of our model. We are aware of the concerns with oversampling as the chances of overfitting are higher and it requires more time to train the data. However, our training accuracy is 96.45% and our testing accuracy is 95.30%. This suggests that our model seems to be learning from the training data without overfitting too much. From this model, we calculated the importance of each variable and found that the most important predictors were patient type (hospitalized or sent home), age, and if the patient had pneumonia or not. From this, we can conclude that these are the features that healthcare professionals should greatly consider when determining if a patient will die from having COVID-19. As well, for people to be aware of if they fall under these categories. In the future, we would hope to analyze more ways to keep the model highly accurate with a low false negative rate and a lower false positive rate. As well as, seeing how we can apply aspects of our model to help predict the risk of death from other infectious disease outbreaks and health crises.

Sources:

https://covid19.who.int/region/amro/country/mx
https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico
https://www.kaggle.com/datasets/meirnizri/covid19-dataset/data