COVID Risk Prediction

SpringBoard Capstone Project

Amanda Fu-Kalilikane

Problem Statement:

The goal of this project is to predict whether a patient will be deemed high risk for COVID-19 based on underlying diseases with a model of 80% accuracy and then find where in Mexico to send more resources.

Background

COVID has ravaged the world for the last two years, quickly becoming one of the most harmful diseases in the world with over 600 million confirmed cases and 65 million deaths. Due to the speed of the spread of this disease, resources have been stretched thin and many issues have arisen in trying to efficiently distribute them.

There are certain groups of patients who are more likely to pass away from complications of COVID 19. People who have underlying medical conditions are more likely to develop serious illness. Understanding what factors would affect whether a patient is high risk or not would aid in effectively providing resources to prioritize those who are high-risk and in need of more focused healthcare.

Client

Organizations that might be interested would include insurance companies to determine policy coverage and pharmaceutical companies involved with the production and distribution of covid vaccines as they may want to prioritize covid vaccines to people considered high risk.

Other organizations or individuals who may be interested would include hospitals and doctors who will be directly involved with the care of patients, as this information would help determine whether a patient is high risk and therefore requiring more care.

Data Wrangling

Source

The data used for this project was first found through Kaggle.

Mexico COVID-19 clinical data | Kaggle

The data originated from the Mexican Government website concerning COVID-19 case information. The following website contains a directory which breaks down and explains the categorical values used for data differentiation. The data is updated daily but data used in this project was most recently updated December 27th, 2022, when the data was first downloaded for use in this project.

Open Data of Mexico - Information regarding COVID-19 cases in Mexico (datos.gob.mx)

Open Data Directorate General of Epidemiology | Ministry of Health | Government | gob.mx (www.gob.mx)

Summary

The database was a single sheet file that contained patient information which included patient demographics, as well as patient covid lab test results, and whether the patient had specific underlying health conditions.

Description

The database was a single sheet file that contained 6,330,966 million rows and 40 columns. After the cleaning process it was reduced to 6,330,961 rows and 30 columns.

Data Cleaning

Values Breakdown

This data had values for all data points, so there weren't any missing values found.

Further breakdown of the data showed

- **Datetime Data** : Date entry and Date symptoms were the dates that patients first exhibited COVID-19 symptoms and when they entered the hospital
    - Date Died had either the date that the patient died or 9999-99-99 if alive
- **Gender Data**: 1 was Male and 2 was Female
- **Entity Data**: The numbers correspond to each of the respective entities of Mexico
- **All Other Categorical Data**:
    - 1 for yes, 2 for no
    - 97 for not applicable, 98 for patient ignored and 99 for unspecified
- **COVID Data**:
    - 1-3 was a positive COVID result
    - 4-7 was a negative COVID result

Missing and Duplicate Values

Given that each row in the data was its own individual registration ID, there were no duplicate values found. If the registration ID was removed then the rest of the data would have some 5000 duplicate values. However there is no way to be sure that these data points are of the same patient under different IDs. it must be assumed that each registration ID is a separate patient.

Since all of the categorical data had a value, these could be considered and treated as missing but they also may be valid data. In this case, the data was treated as not missing data.

Columns

The data consisted of patient demographic information, such as age, gender and entity of residence as well as medical information, including whether the patient had specific medical conditions or if they had been tested for COVID-19 using an antigen test and if these had come back.

There were many columns that seemed unnecessary to the problem statement and were dropped, some examples include

- "Updated Date" - since it was the same for all rows
- "Registration ID" - since it was different for every row
- Some demographic information such as as "Migrant", "Indigenous" or "Language" which explained whether the patient came from a different country or not
- Some Medical information including "Lab Sample taken", "Lab Sample Result", because the only feature of interest at this time is whether the patient had COVID or not.

Exploratory Data Analysis

Chi Square Test

First a Chi Square test was performed on the data to ensure that there may be a relationship between the features and the target features (COVID positive or not), and found that all features had a P-value below 0.05 showing that the null hypothesis would be rejected and all features might have a relationship with the target variable.

Death and COVID Diagnosis

Further Exploration into data concerning COVID and death of the patients showed that 49.58% of the patients in this dataset tested positive for COVID but less than 1% of patients died. When comparing the numbers of patients who died, 25188 died with a positive COVID result and 14805 died with a negative COVID result showing that 60% of the patients who died had a positive COVID result.

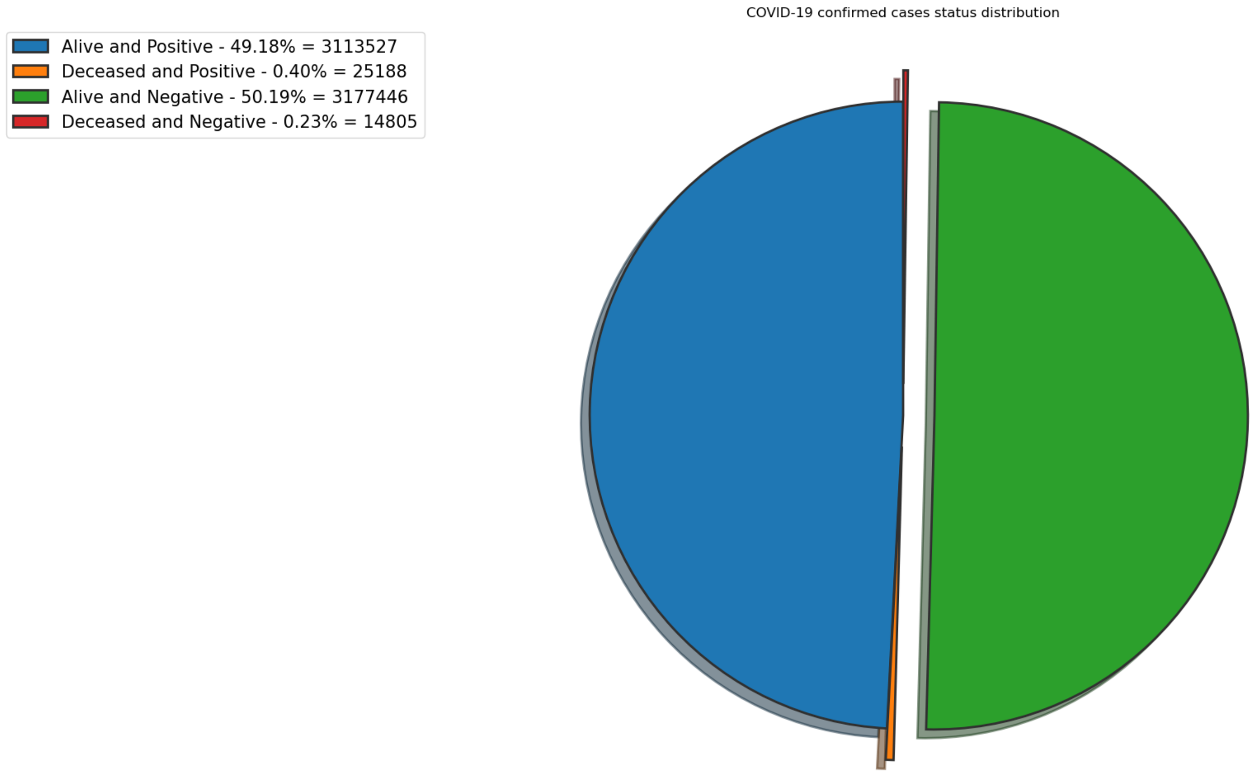


Figure 1: Pie Chart depicting patient percentages

Cases and Timeline

The timeline of cases was explored. There was a trend seen that the number of deaths peaked around the same time that the number of cases would peak, this makes sense as with a higher number of cases of patients entering the medical care facilities, so would the number of patients dying.

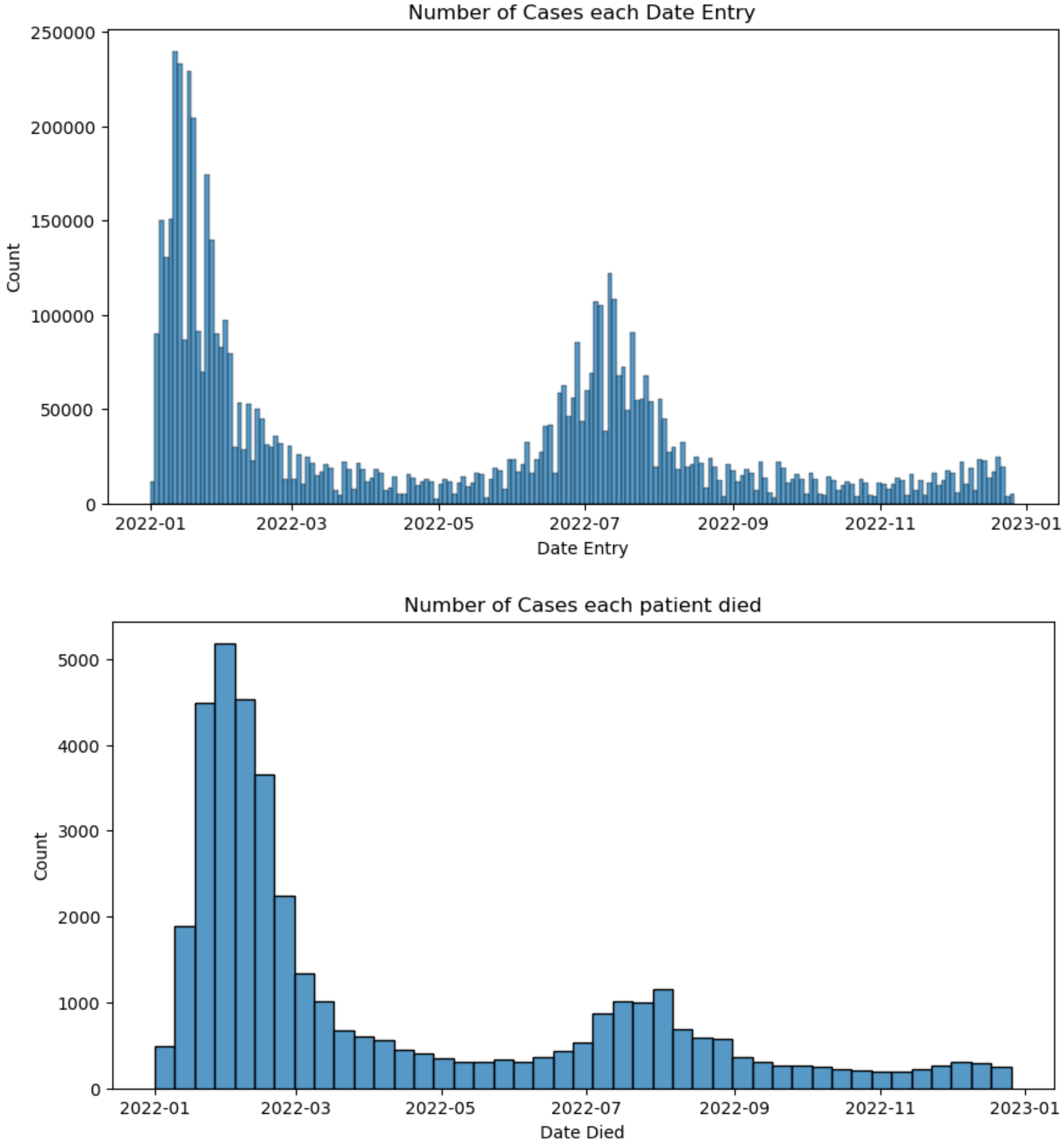


Figure 2-3: Histogram depicting number of patients entry and when they died.

Age Distribution

There were some outliers in the data at this time with patients with ages going as high as 266. Because of patient confidentiality there is no way to check and find the correct age for these patients. With further research, there have been Mexican residents who reach ages as high as 120 years of age, so any ages above 120 were deemed incorrectly labeled and dropped.

The age distribution was found to be very widely distributed with the mean of all patients around 37.3 years.

The mean age of the patients who are alive is 37.1 years.

The mean age of the patients who died is 66.1 years.

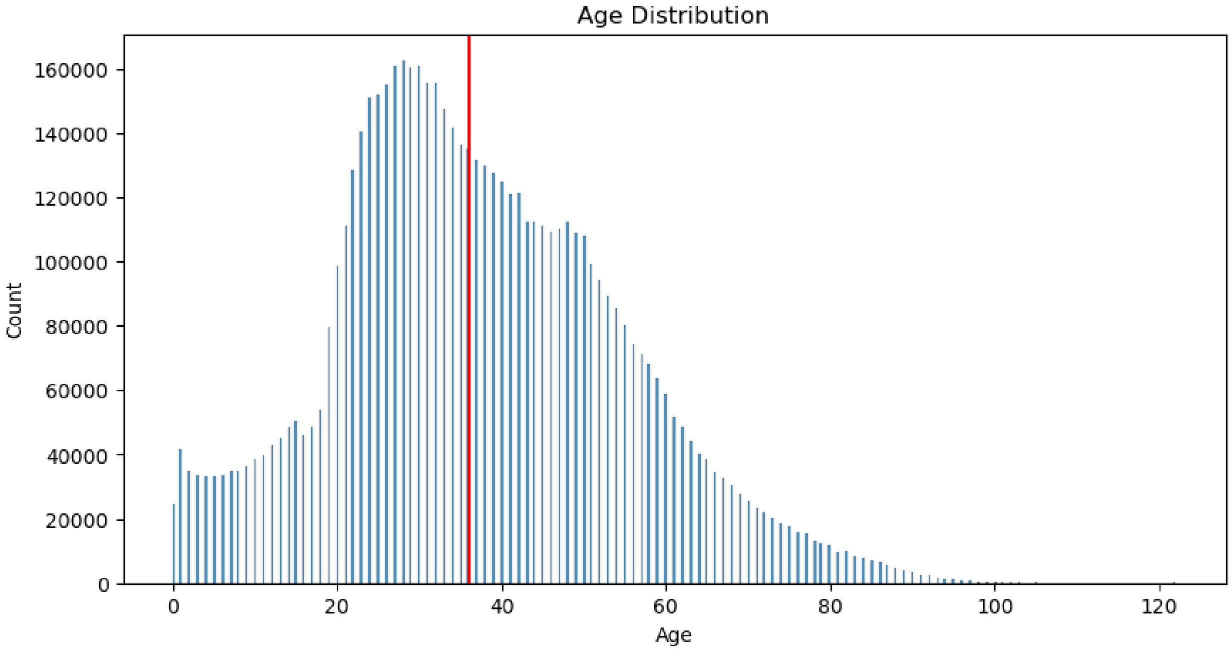


Figure 4: Histogram of Age distribution

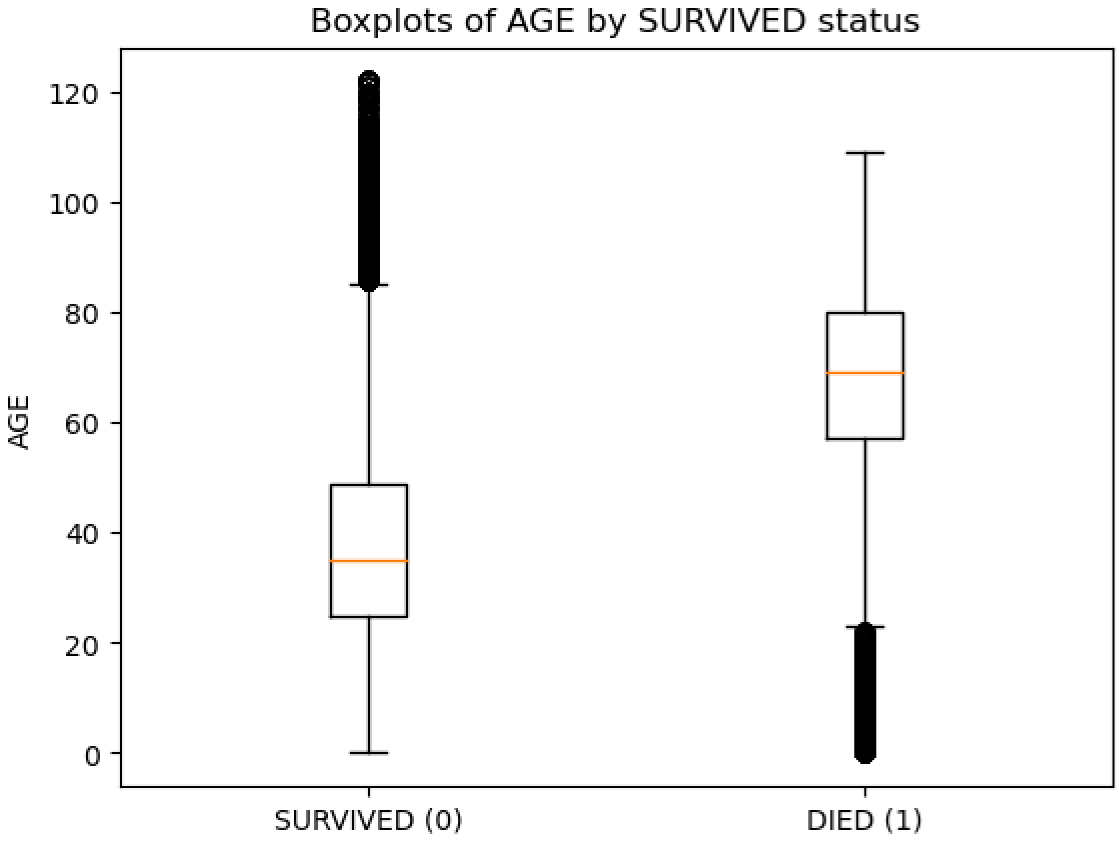
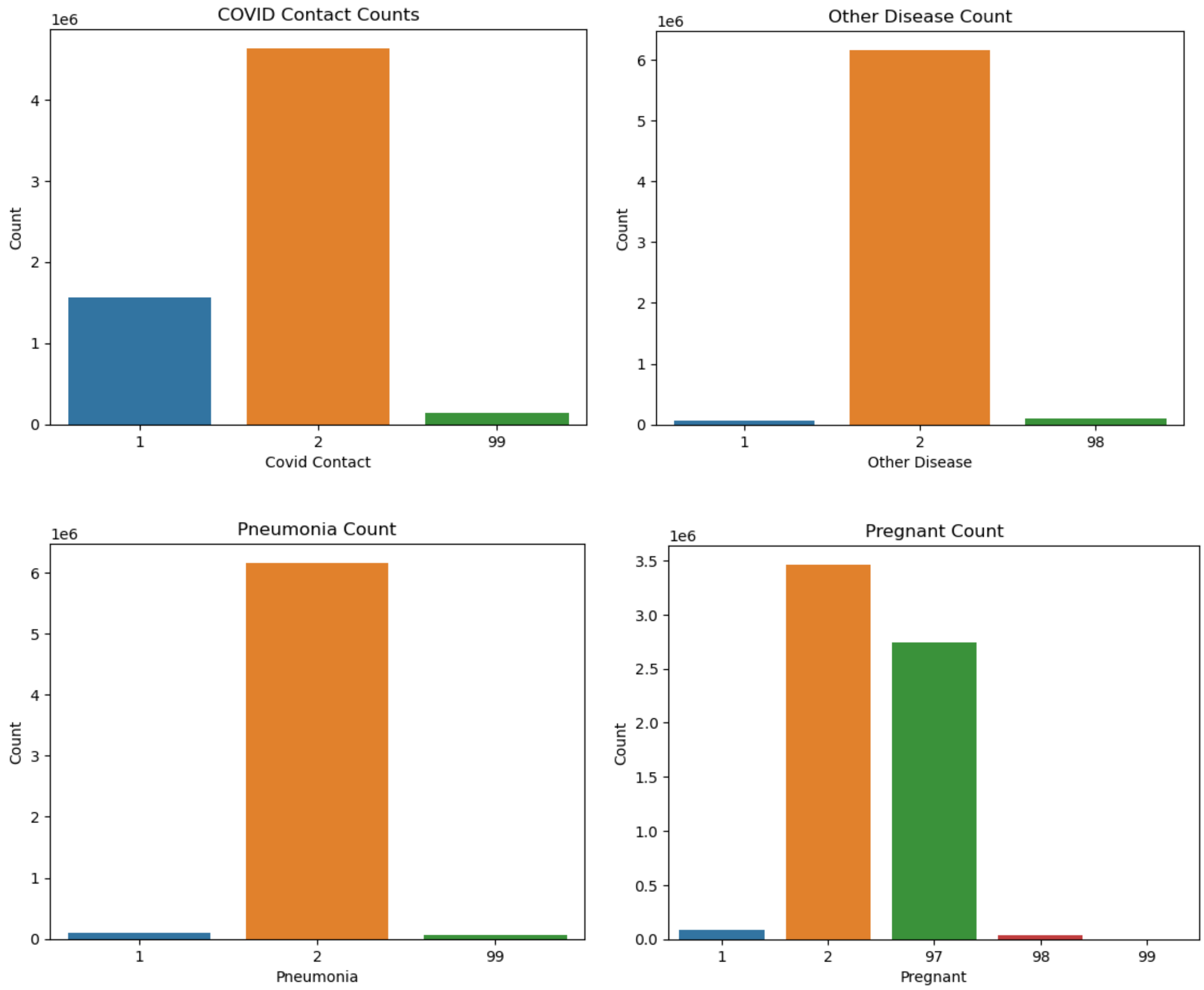


Figure 5: Box plot of age distribution based off of patients who died or survived

Other Features

Other features were explored but did not show any distinct trends. Many of the given categorical data, the patients answered no or the particular features, such as pregnancy with a male, was inapplicable.

COVID Contact and pneumonia were found through feature selection to be the most important in creating an accurate predictive model.



Figures 6-11: Bar chart of various features

Entity Breakdown

After determining whether a patient is considered high risk, the goal of this project will be to find where the patient is located in order to provide resources to the hospitals and medical care facilities that would need it the most.

Looking at the overall patient spread, the entities that have the most patients would be the City of Mexico, the state of Mexico and the entity of Nuevo Leon.

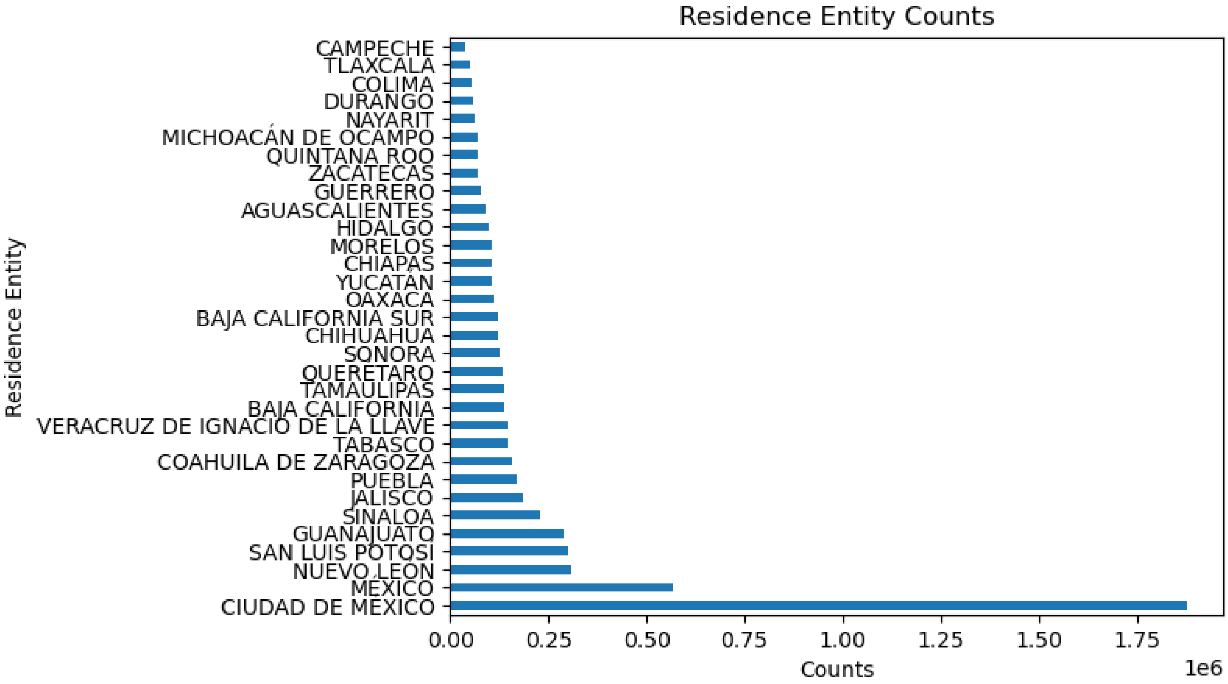


Figure 12: Mexican Entity COVID Cases

Model Selection

Feature Selection

Used feature selection to choose the top ten features for modeling

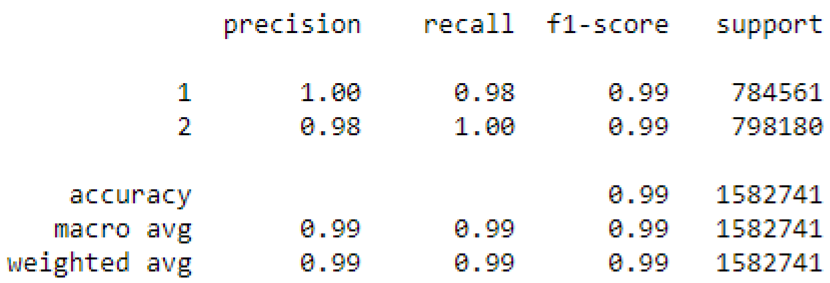
- Entity
- Intubated
- Pneumonia
- Age
- Pregnant
- Other Disease
- COVID Contact
- Sample Result
- Antigen Result
- ICU

Models

For this classification problem, the following Models were used

- Random Forest Classifier
    - Accuracy of 98.87%
- Logistic Regression
    - Accuracy of 57.24%
- Gradient Boosting Classifier
    - Accuracy of 98.82%

Both Random Forest Classifier and Gradient Boosting Classifier performed extremely well in this project, with Random Forest Classifier just slightly better.



```
              precision    recall  f1-score   support

           1       1.00      0.98      0.99    784561
           2       0.98      1.00      0.99    798180

    accuracy                           0.99   1582741
   macro avg       0.99      0.99      0.99   1582741
weighted avg       0.99      0.99      0.99   1582741
```

Figure 13: Metric Scores for Random Forest Classification

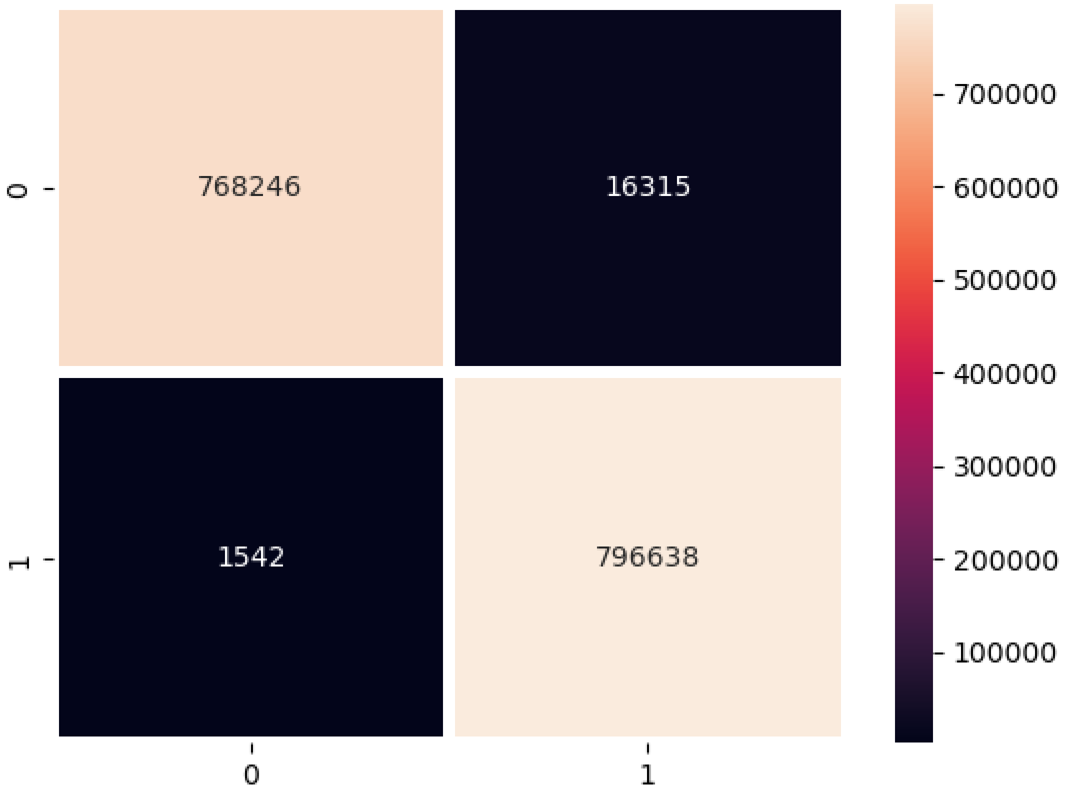
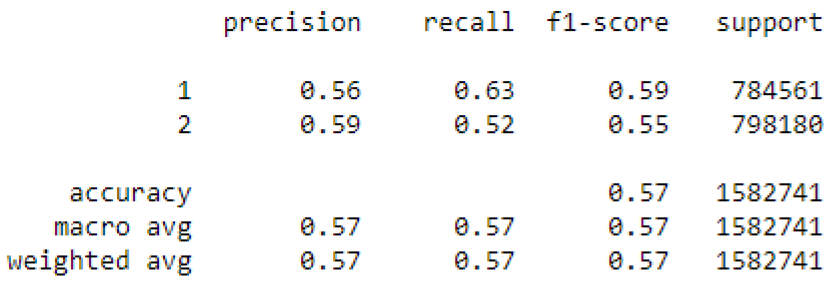


Figure 14:Confusion Matrix for Random Forest Classification



```
              precision    recall  f1-score   support

           1       0.56      0.63      0.59    784561
           2       0.59      0.52      0.55    798180

    accuracy                           0.57   1582741
   macro avg       0.57      0.57      0.57   1582741
weighted avg       0.57      0.57      0.57   1582741
```

Figure 15: Metric Scores for Logistic Regression

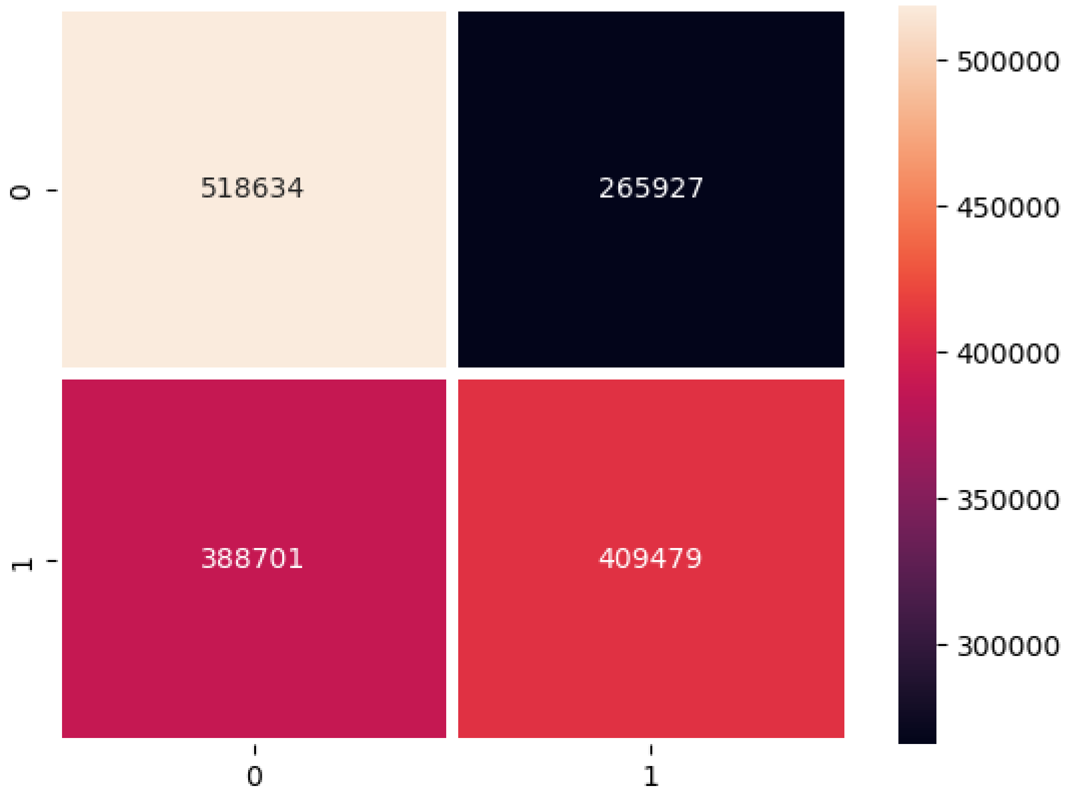
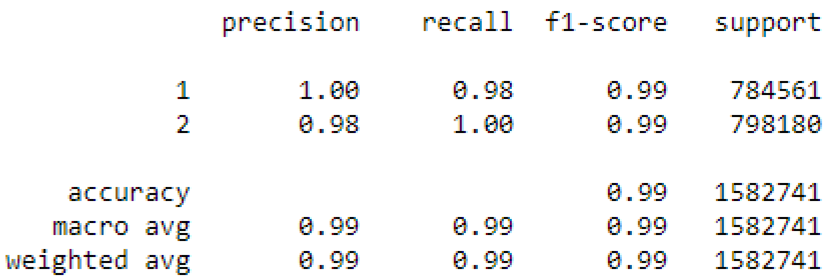


Figure 16: Confusion Matrix for Logistic Regression



```
              precision    recall  f1-score   support

           1       1.00      0.98      0.99    784561
           2       0.98      1.00      0.99    798180

    accuracy                           0.99   1582741
   macro avg       0.99      0.99      0.99   1582741
weighted avg       0.99      0.99      0.99   1582741
```

Figure 17: Metric Scores for Gradient Boosting Classification

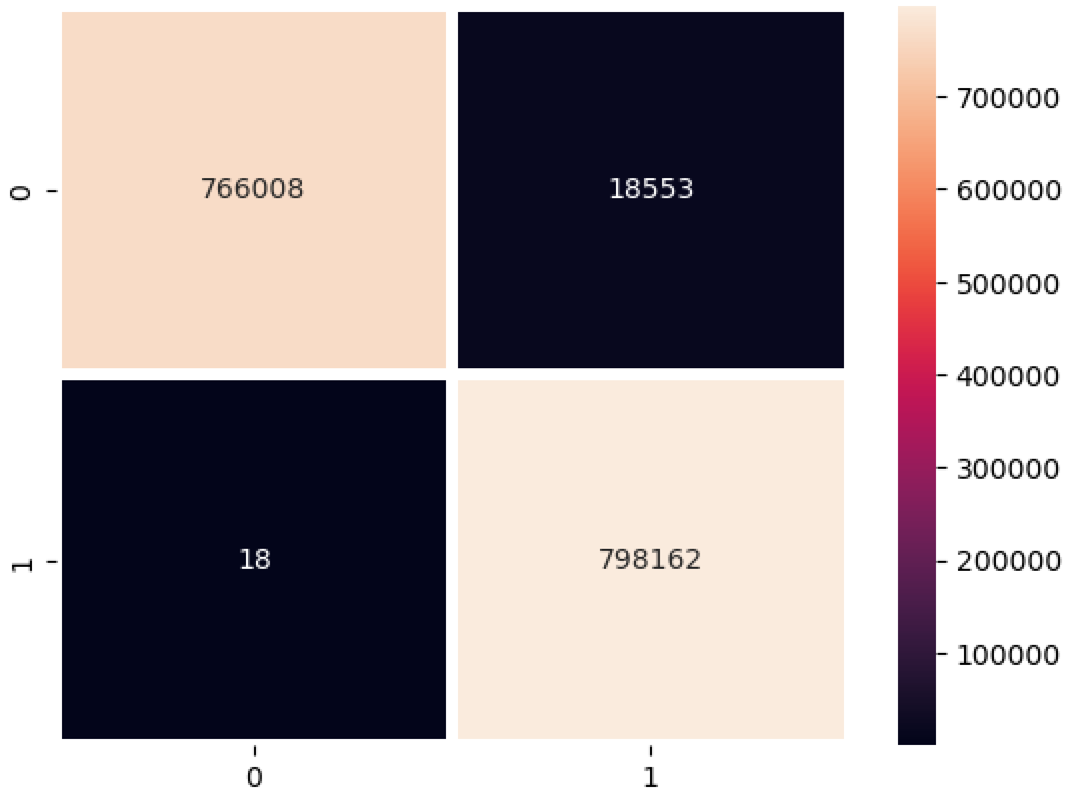


Figure 18: Confusion Matrix for Gradient Boosting Classification

Conclusions

The feature that seemed most determinant on whether a patient would be contracting COVID would be whether they had contact with another person with COVID. All other health features were not largely considered determinant on whether the patient contracted COVID or if they would die from COVID. As it was consistently approximately half of the given population of patients who would contract COVID. It also seems that the patients seem to mostly be from the City of Mexico which would be the first choice to send resources to.

Future Considerations

The focus on this specific project was to determine whether a patient would be considered high risk for COVID in the year of 2022. Given that the spread of COVID had begun to slow down greatly during this time, due to vaccinations and previous infections of COVID. It would be very interesting to do a similar project with previous years of COVID information and compare.