COVID RISK PREDICTIONS

Amanda Fu-Kalilikane

Problem Statement

How to predict where in Mexico to send more resources due to the higher population of patients that are deemed high risk for COVID-19 based on underlying diseases with a model of 80% accuracy.

Solution

Although there are many health features that would deem a person high risk in general.

This project showed that the features that deem the highest risk would be where you are located, whether you have had COVID contact and the health feature of pneumonia.

Data Cleaning

Understanding the Data

- **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

Data Cleaning

No Missing or duplicated data due to individual registration IDs

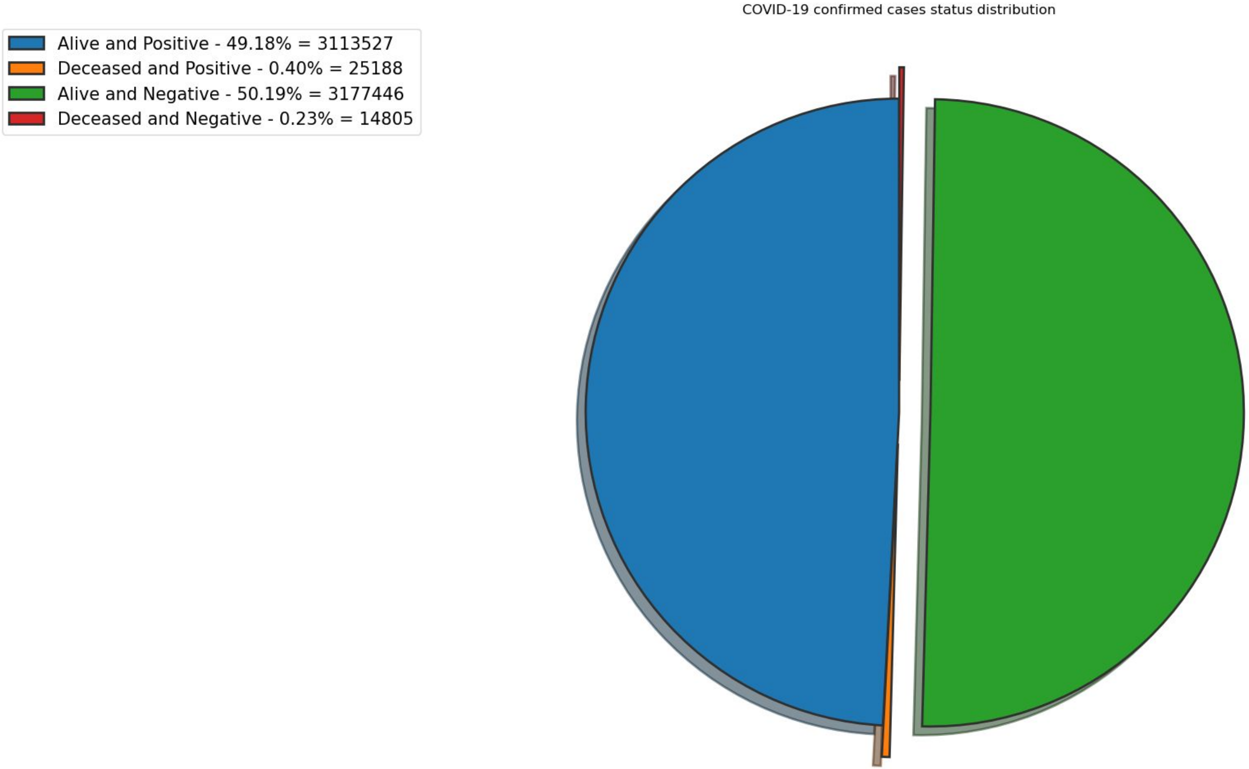
There were many columns that seemed unnecessary to the problem statement and were dropped, some examples include

- “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.

Data Exploration: Deaths

About 50% of the patients included in the data was tested positive for COVID -19

Less than 1% of the total patients died with about 60% of those patients being COVID positive

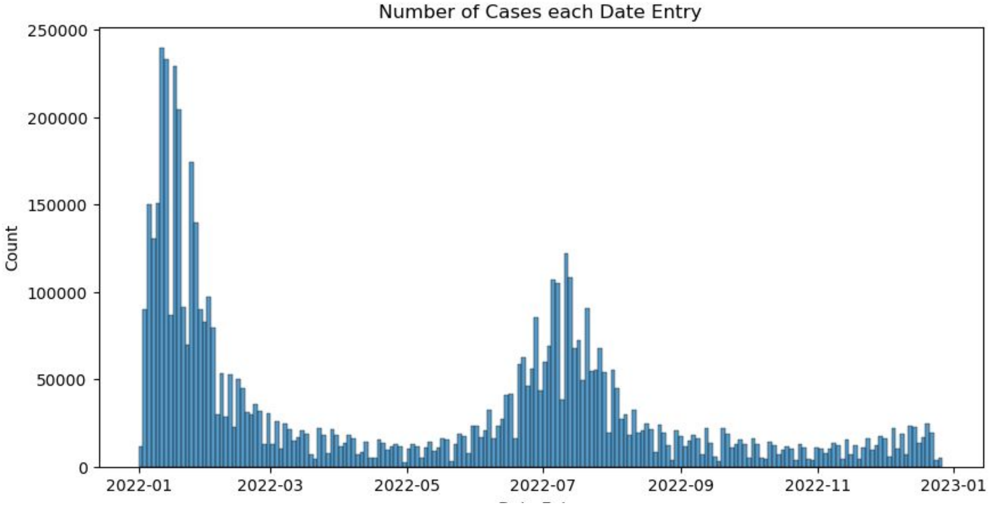


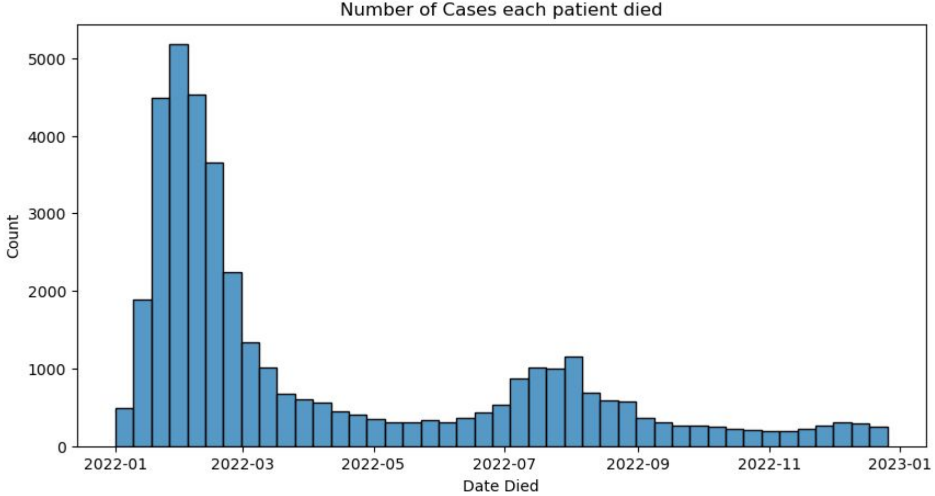
COVID-19 confirmed cases status distribution

Data Exploration: Timeline

Exploring the datetime

Patient deaths followed the trend of patient entry, however at a much smaller scale



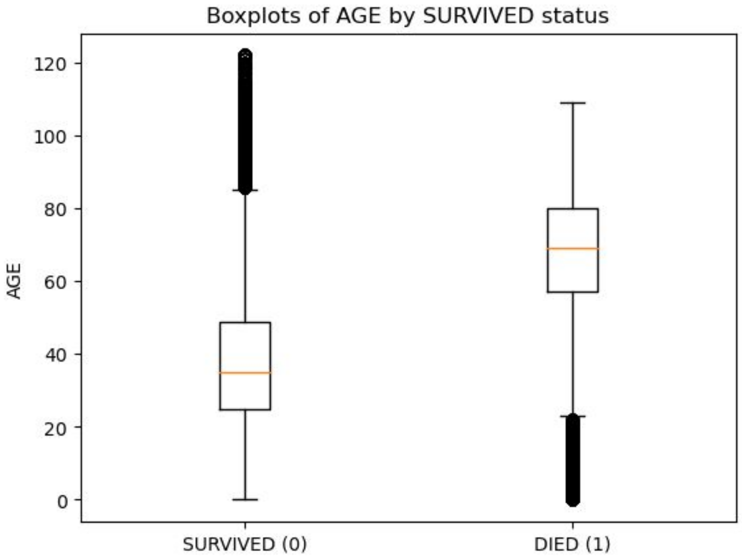


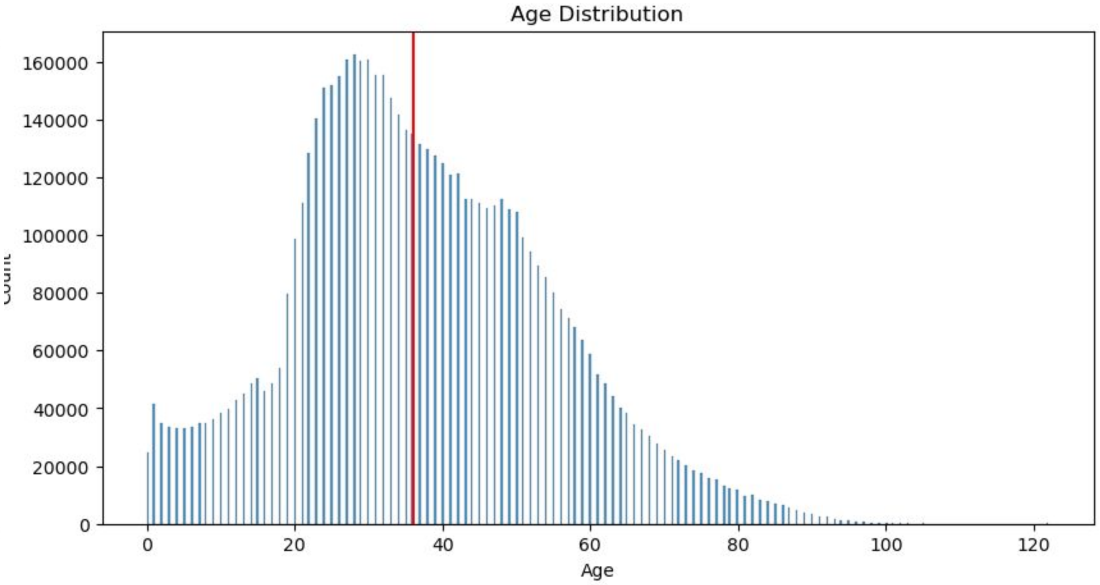
Data Exploration: Age Distribution

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.

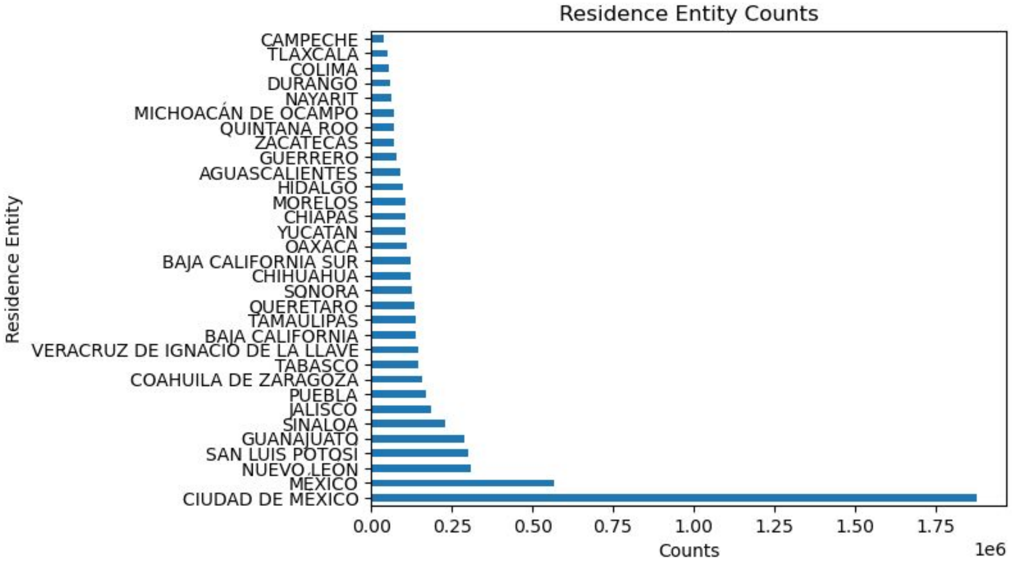
The mean age of the patients who died is 66.1 years.





Data Exploration: Entity Breakdown

In exploring the data involving the entities of Mexico, it is found that a large amount of the patients reside in the City of Mexico



Feature Selection

Used feature selection to choose the top ten features for modeling

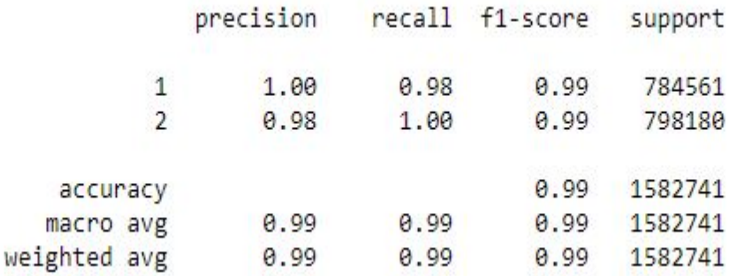
- Entity
- Intubated
- Pneumonia
- Age
- Pregnant
- Other Disease
- COVID Contact
- Sample Result
- Antigen Result
- ICU

Modeling:

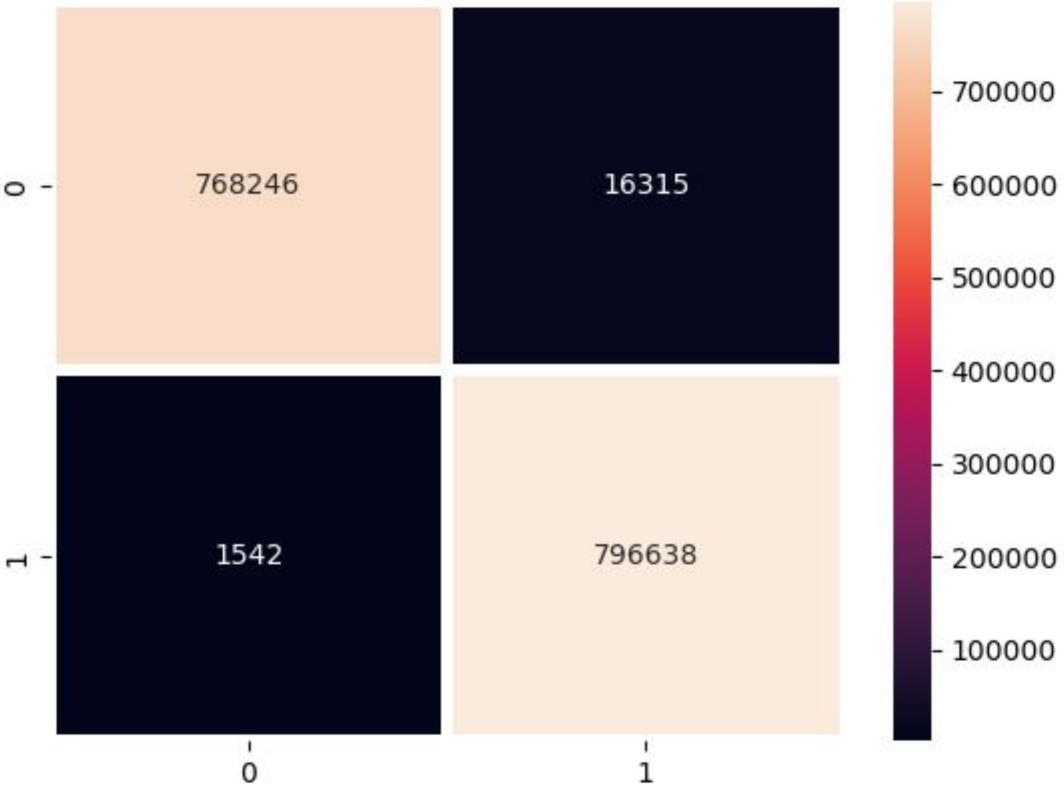
Random Forest Classifications

With an accuracy of 98.87%

Random Forest Classifier was the highest performing Model



```
              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
```

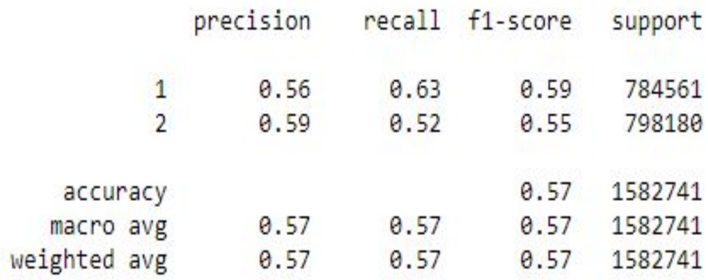


Modeling:

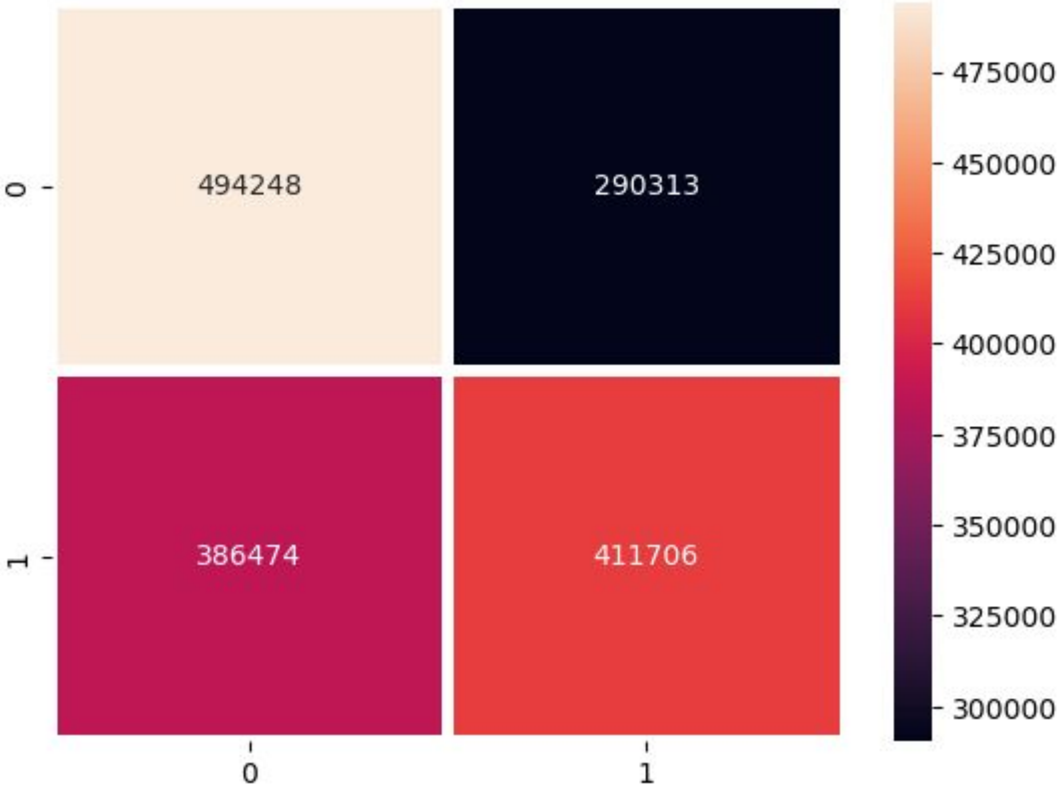
Logistic Regression

With an accuracy of 57.24%

Logistic Regression was the lowest performing Model



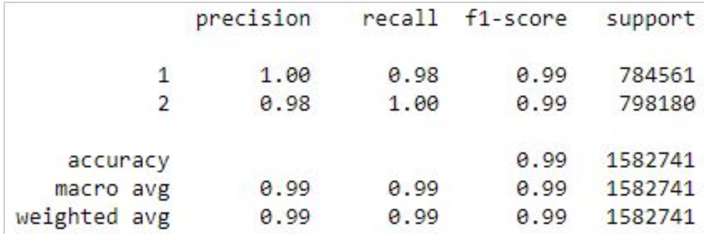
```
              precision    recall  f1-score   support

           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
```



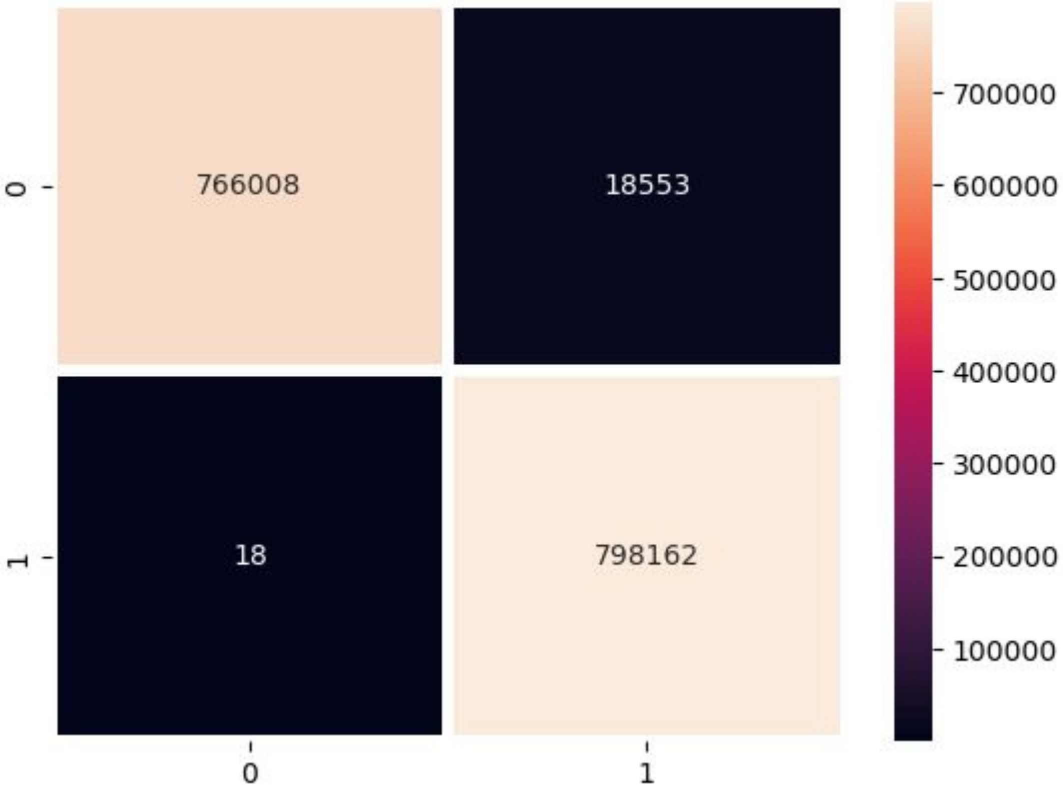
Modeling:
Gradient Boosting

With an accuracy of 98.83%

Gradient Boosting was just slightly lower than the Random Forest Classifier Model but still performed extremely well



```
              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
```



Future Considerations

This project was specifically of COVID in Mexico during 2022

Other future projects to consider could include

- Comparing this project with COVID in Mexico in 2021 or 2020
- Comparing other countries based off of their own COVID responses and restrictions