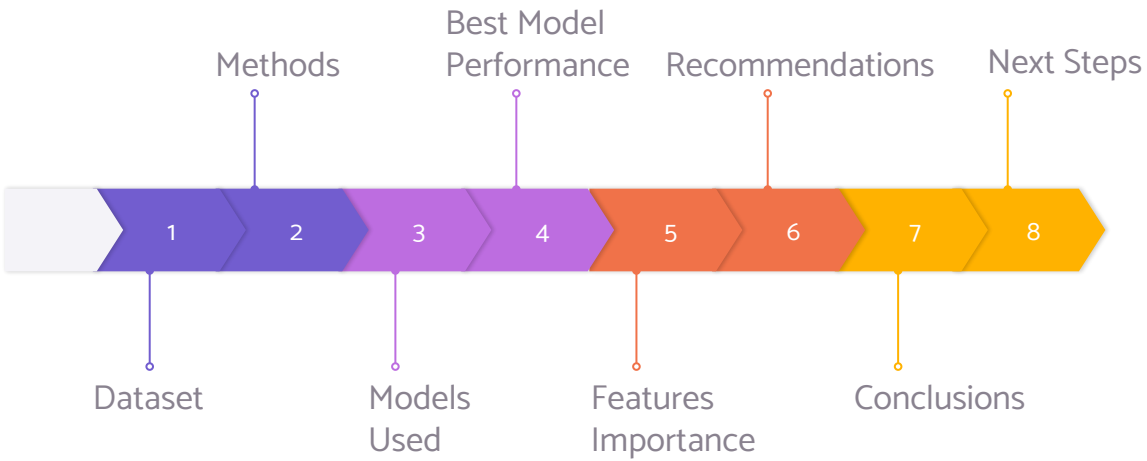
Covid-19 Hospitalization Predictor

Business Problem

- During the Covid-19 pandemic one of the major issues was the shortage of medical resources and a system to distribute them.
- In this project we built a model that can predict whether a patient is at risk to be hospitalized, to help estimate the amount of people that will require hospitalization for the next pandemic.
- We also found the main factors that put a patient at risk for hospitalization, to help the CDC target those factors in the population, to bring down hospitalization rate.

Roadmap


Methods
Best Model Performance
Recommendations
Next Steps
1
2
3
4
5
6
7
8
Dataset
Models Used
Features Importance
Conclusions

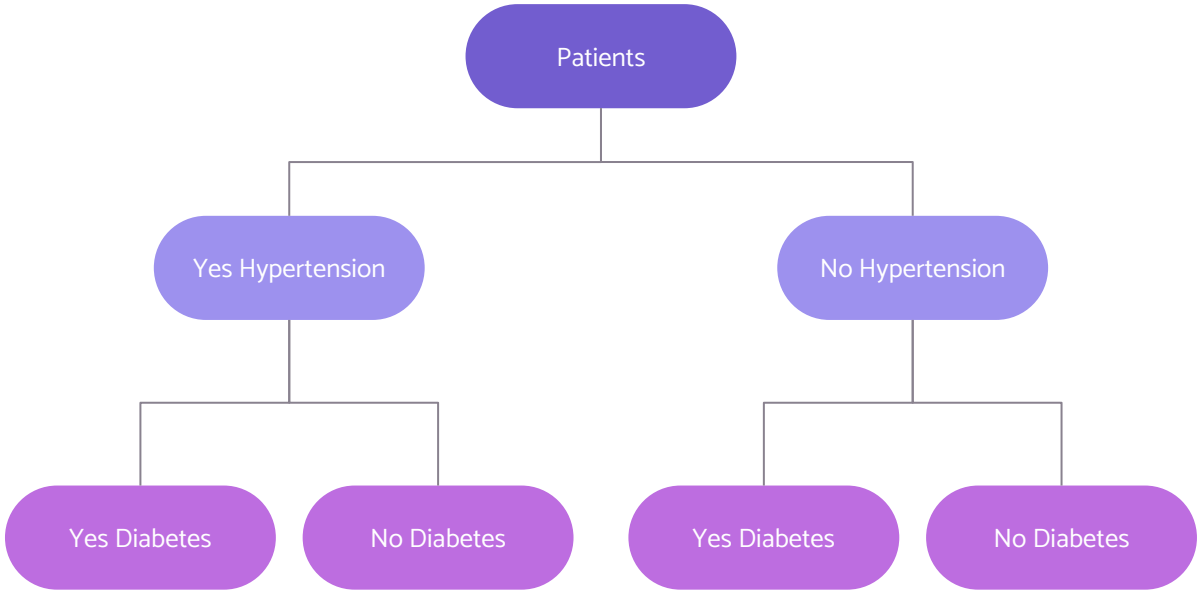
DataSet:

- dataset from kaggle, provided by the Mexican government
- contains 21 unique features with information about 1,048,576 patients
- 392 thousand Covid positive patients
- data collected between January 2020 and May 2021
- contains info about pre-conditions and hospitalization

Methods:

- *Data Cleaning*
- *Logistic Regression Model*
- *Decision Tree Classifier*
- *Random Forest Classifier*
- *Tuning of Random Forest*
- *Gradient Boost*
- *Study of most Relevant Features*

Decision Tree Classifiers



Models Used:

Decision Trees

Decision Trees divide the sample repeatedly, based on the factor that gives the most information, to reach a conclusion about classification

Random Forests

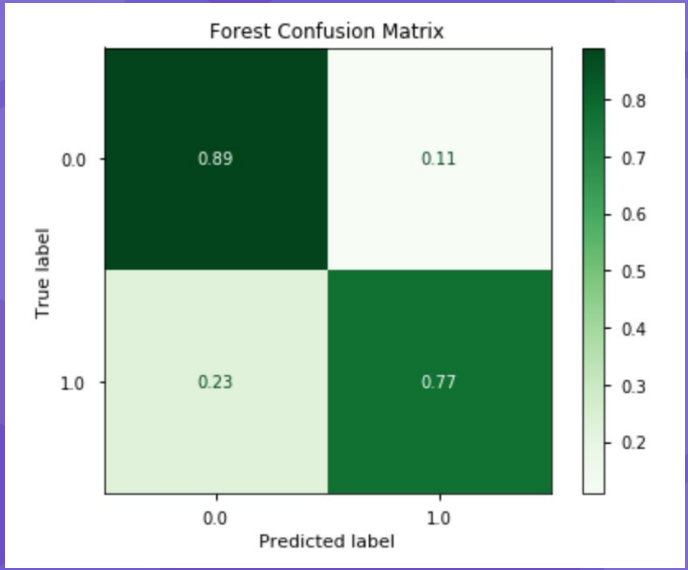
Use several different trees to produce a more precise model that adapts better to unseen data

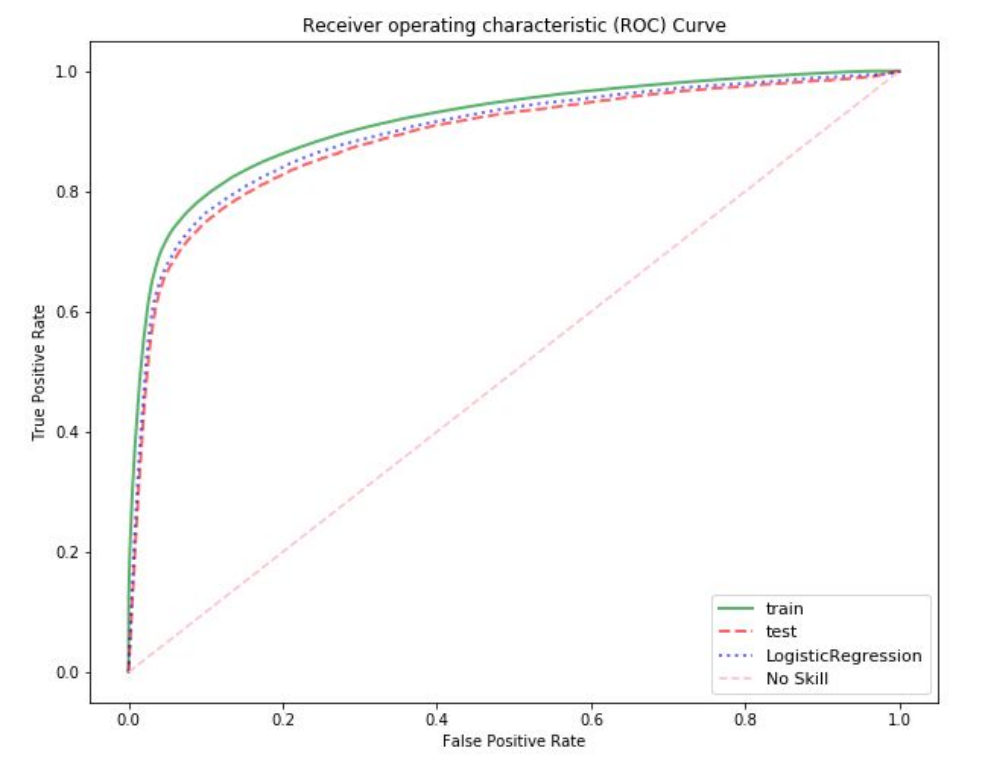
Gradient Boost

Starts with simple Decision Trees and improves them learning from the mistakes of the previous trees

Results

Best Model: Random Forest

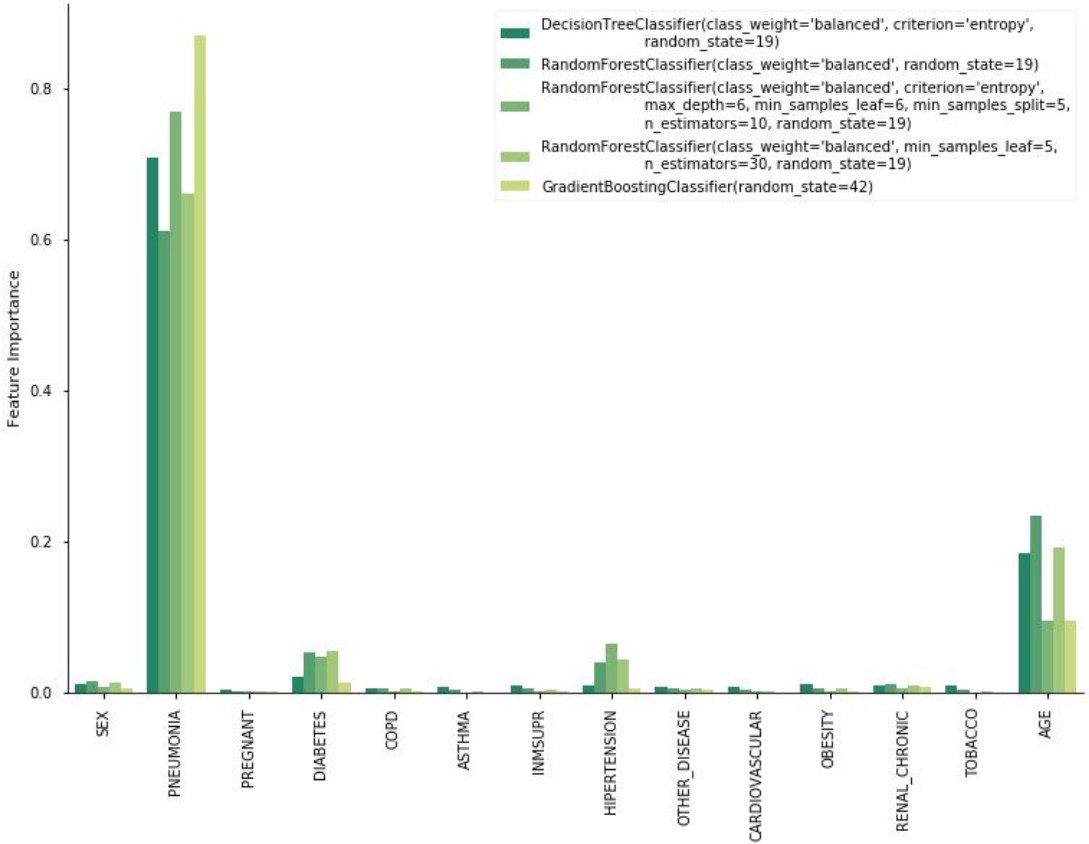


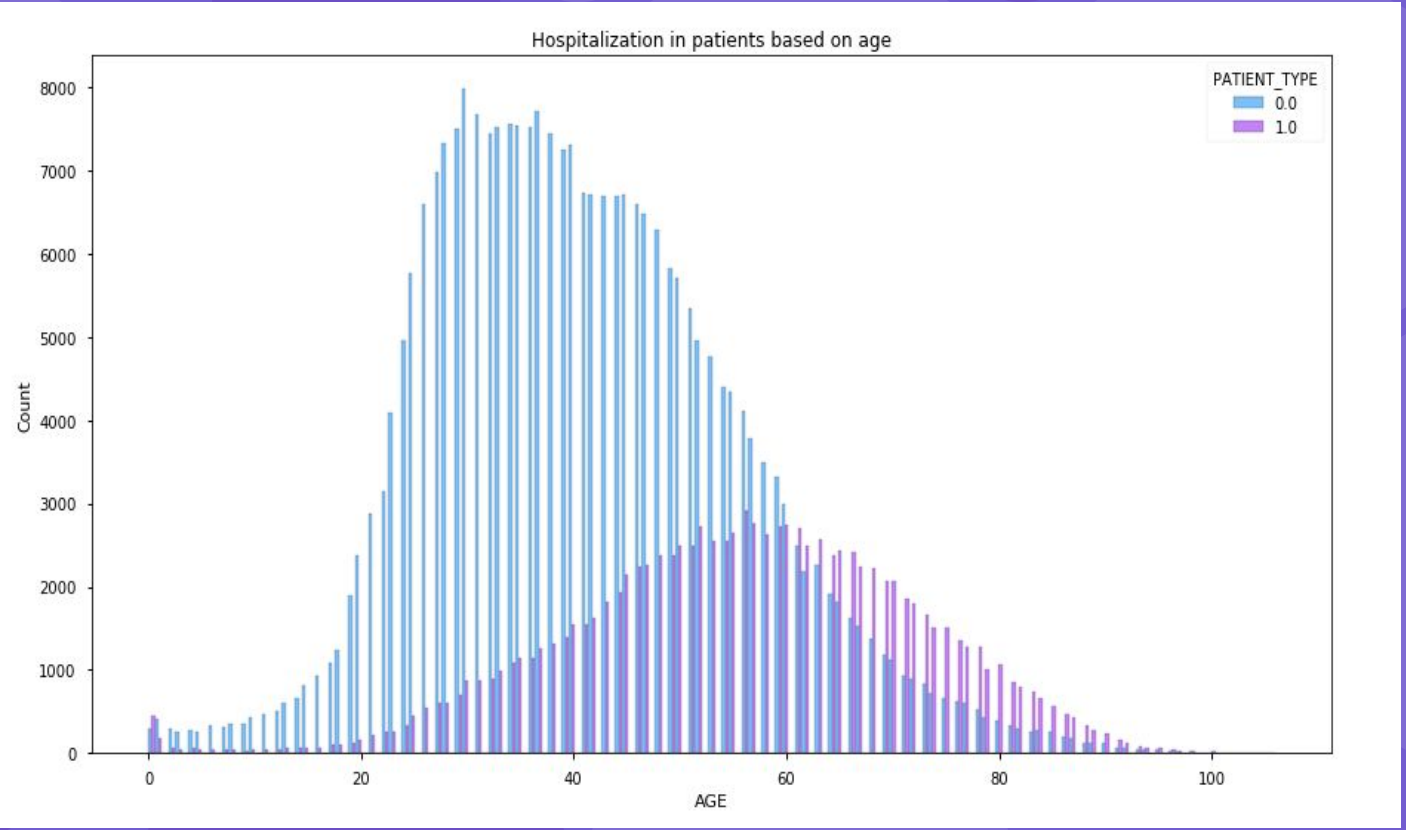


Recall 77% – F1 score 76% – AUC score 90%

Most relevant features

1. Pneumonia
2. Age
3. Hypertension
4. Diabetes
5. Sex





Age

Hospitalization rate by age

8% for individuals between 20 and 30 years old

~ 60% for individuals between 60 and 70 years old

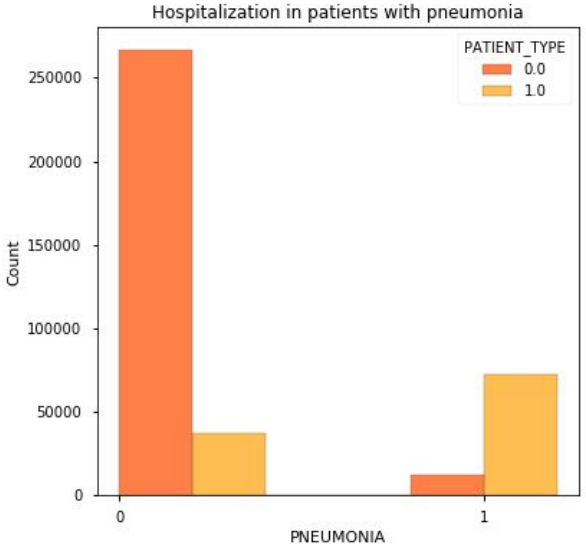
~70% for individuals between 70 and 80 years old

61% for infants less than 1

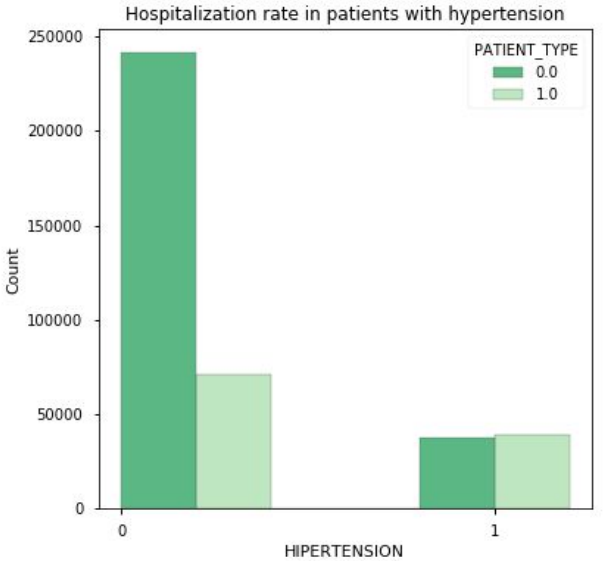
30% for kids between 1 and 2

18% for kids between 2 and 3

Study of the features

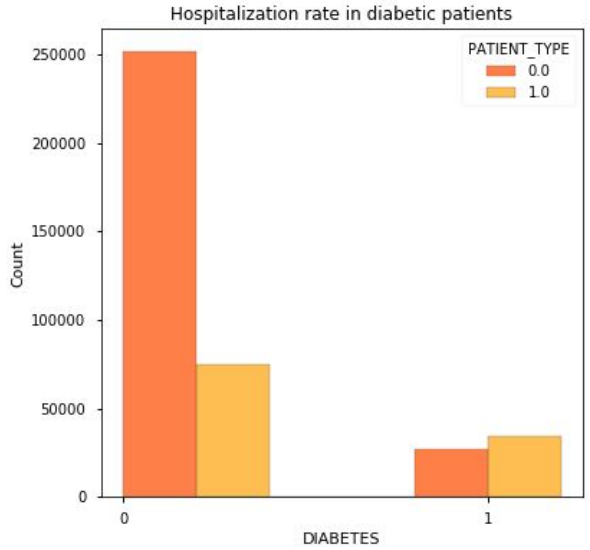


86% of patients with pneumonia were hospitalized
12% of patients without pneumonia were hospitalized

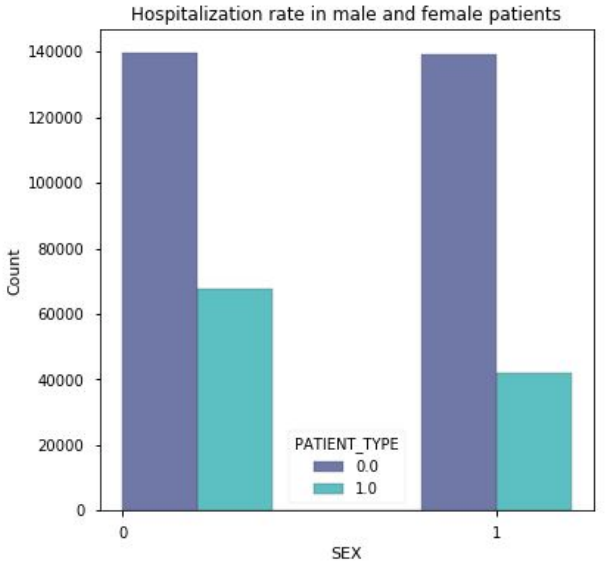


51% of patients with hypertension were hospitalized
23% of patients without hypertension were hospitalized

Study of the features



56% of patients with diabetes were hospitalized
23% of patients without diabetes were hospitalized



Hospitalization based on sex
23% of females were hospitalized
33% of males were hospitalized

RECOMMENDATIONS

Treat

Lauch health campaigns to treat and prevent pneumonia, hypertension and diabetes

Inform

Inform the population of the high-risk factors so that groups of people that are considered at risk can take extra precautions

Protect

Have healthcare providers treat and target these conditions specifically and follow high-risk patients more closely

Conclusions

Based on our model we predicted the number of patients that needed hospitalization, and found the factors that put patients more at risk.

When the next pandemic hits, the same model could be updated based on the first patients that get infected, and determine risk factors much quicker.

Applying the model in a specific geographic area, the expected rate of hospitalization could be calculated, so that hospitals can be better prepared to treat all the patients in need.

Next Steps

To improve our model we could:

- Use a broader sample of patients
- Study the effect of multiple factors' interaction
- Refine the model by area to account for differences in population
- Automate the process of feature extraction

THANKS!

Any questions?

You can find me at:

marianlkuzmin@gmail.com