Analysis of COVID-19 Patients with Pre-Existing Conditions

Alex Bass, Connie Cui, Peumali Surani Withanage, Seth Galluzzi

Aim

The aim of this project is to gain an understanding of how pre-existing medical conditions impact COVID-19 patients in hospitals.

Rationale

We will accomplish this by comparing the mortality rates of patients with pre-existing conditions to the mortality rates of patients without pre-existing conditions.

We will then use our data to calculate the probability of survival for a new patient.

The Data

The data set we are exploring was released by the Mexican government and contains information on hospitalized COVID-19 patients.

The data includes variables such as age, sex, diabetes, pneumonia, and obesity.

Most of the data was categorical with 0 and 1 values.

	Unnamed: 0	sex	patient_type	pneumonia	age	diabetes	copd	asthma	inmsupr	hypertension	other_disease	cardiovascular	obesity	renal_chronic	tobacco	contact_other_covid	covid_res	died
0	0	0.0	1.0	0.0	27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
1	1	1.0	1.0	0.0	56	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0
2	2	1.0	1.0	0.0	34	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
3	3	1.0	1.0	0.0	34	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0
4	4	1.0	1.0	0.0	49	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0

Data Cleaning

Though our data was relatively clean, we did complete some cleanup before creating visualizations. For example, we dropped an unnamed column from the data set.

We also split the data into training and test sets to test our analysis and run our data through machine learning classifiers to determine the best model to make our predictions.

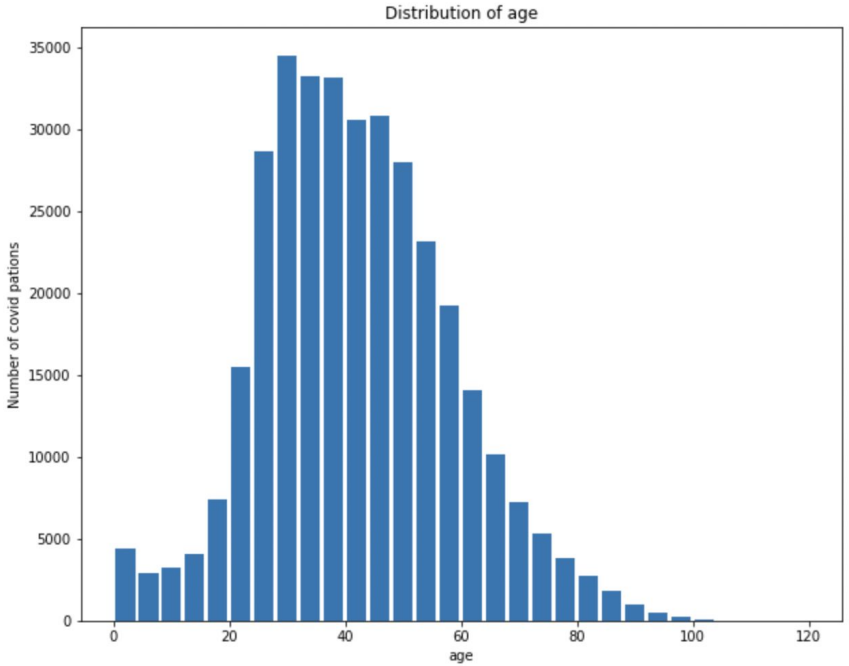
	sex	patient_type	pneumonia	age	diabetes	copd	asthma	inmsupr	hypertension	other_disease	cardiovascular	obesity	renal_chronic	tobacco	contact_other_covid	covid_res
0	0.0	1.0	0.0	27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
1	1.0	1.0	0.0	56	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0
2	1.0	1.0	0.0	34	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0
3	1.0	1.0	0.0	34	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
4	1.0	1.0	0.0	49	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0

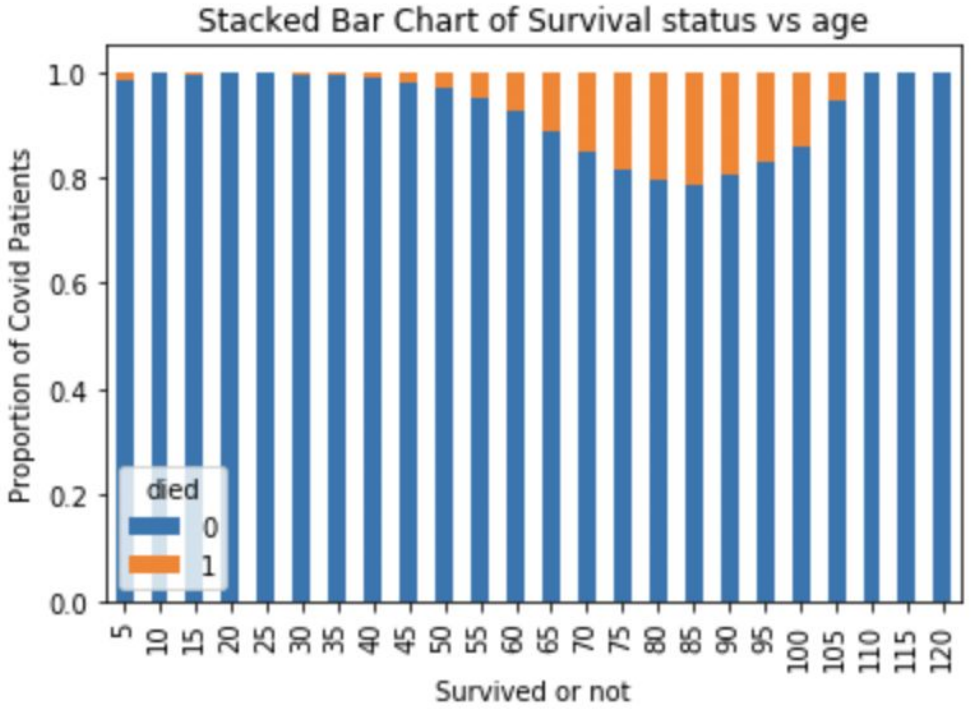
```
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    train_size=0.7,
                                                    random_state=5100)
```

Identifying Key Visualizations

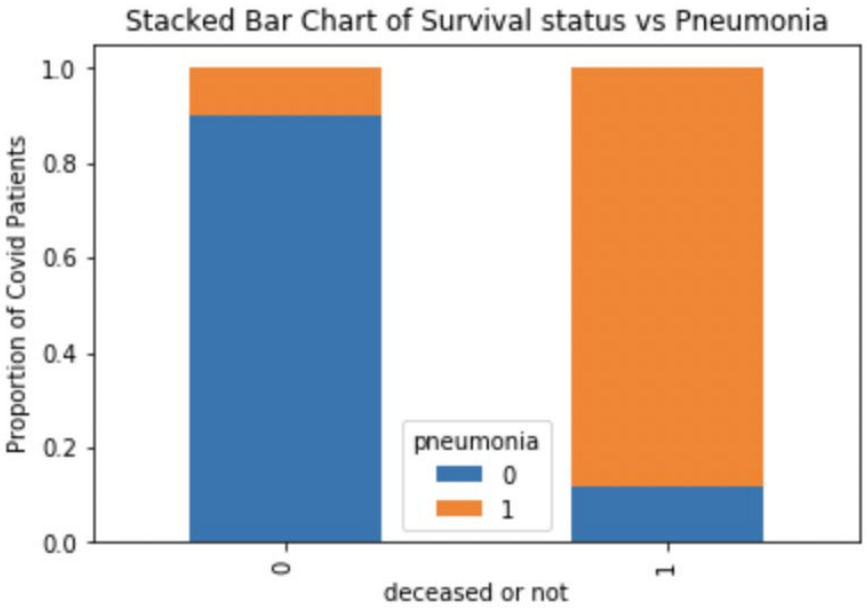
- We used histograms, bar charts, and proportions to interpret our data effectively.
- Each of our visualizations was an opportunity to further our understanding of the data.

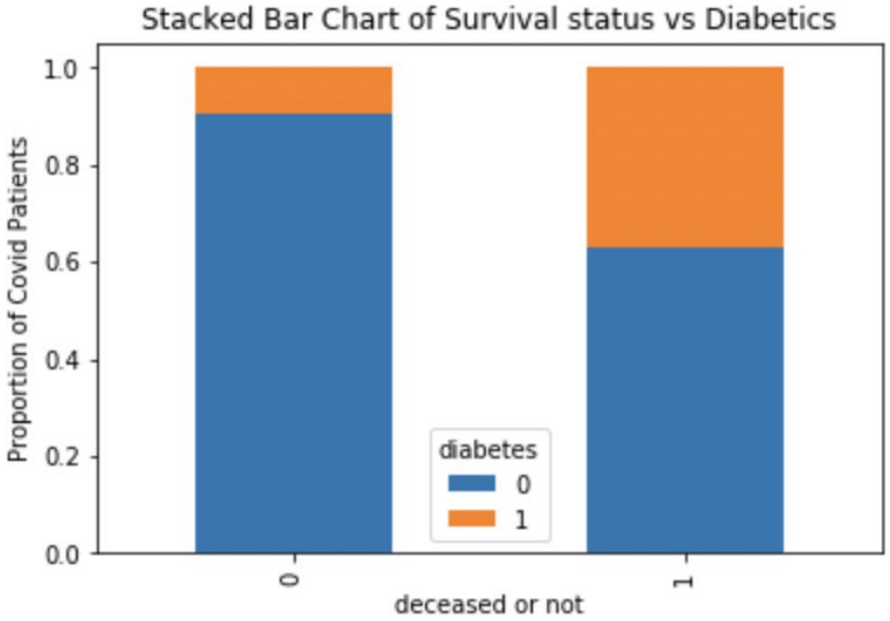
Identifying Key Visualizations





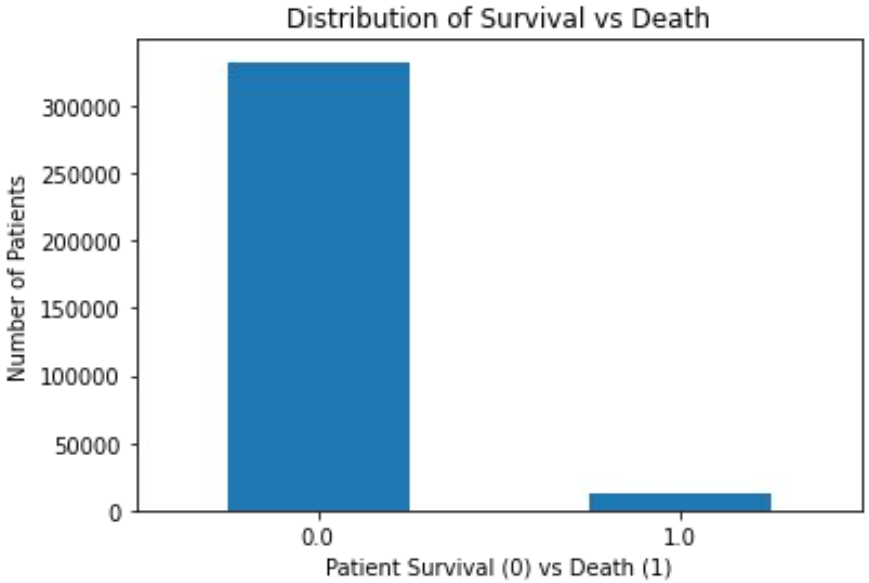
Identifying Key Visualizations





Knowledge from Visualizations

- Fortunately, one major issue with our data was the large number of survivors!
- To prevent Type II error, we created a balanced dataset with oversampling techniques.

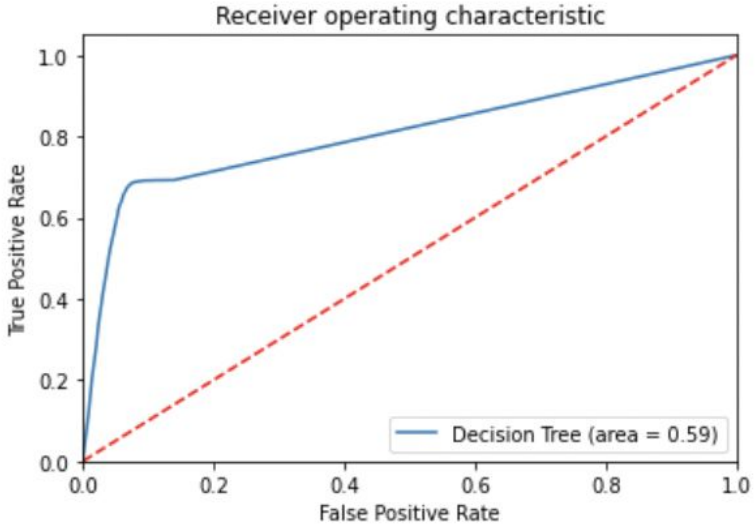


Model Selection

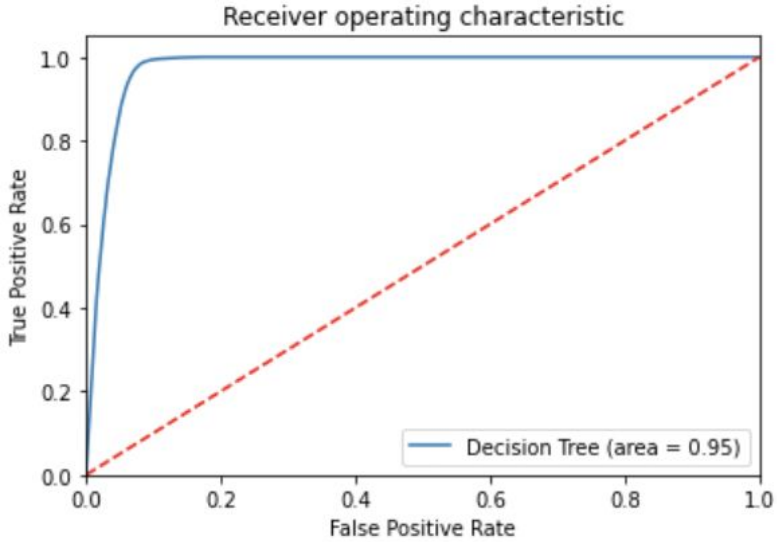
- We used machine learning to train three models with the original, unbalanced data and three models with balanced data that we created with oversampling.
- Classifiers used:
  - Logistic Regression
  - Decision Tree
  - Random Forest

The Impact of Balancing Our Data

```
Accuracy: 0.955850336204073
Precision: 0.36511522819701764
Recall: 0.2027094831911691
F1 Score: 0.26068720761413133
AUC: 0.5943170737935408
```




```
Accuracy: 0.9512115701568503
Precision: 0.9241980861334061
Recall: 0.9831538276745676
F1 Score: 0.9527648037404028
AUC: 0.9511809774999959
```



Model Selection Results

We used machine learning to train three models with the original, unbalanced data and three models with balanced data that we created with oversampling.

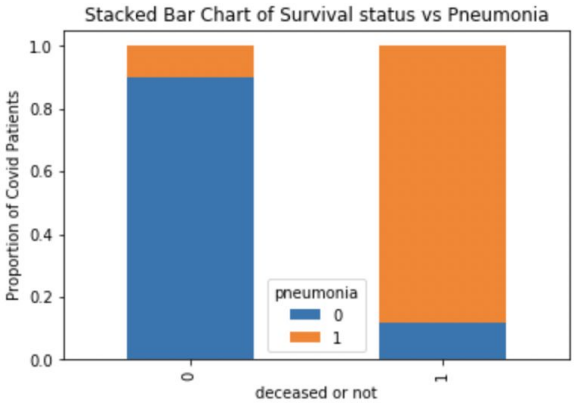
Original Dataset	**Logistic Regression**	**Decision Tree**	**Random Forest**
Accuracy	0.9613	0.9559	0.9569
F1 Score	0.2424	0.2607	0.2769

*F1 Score = weighted average of Precision and Recall (takes both false positives and false negatives into account)

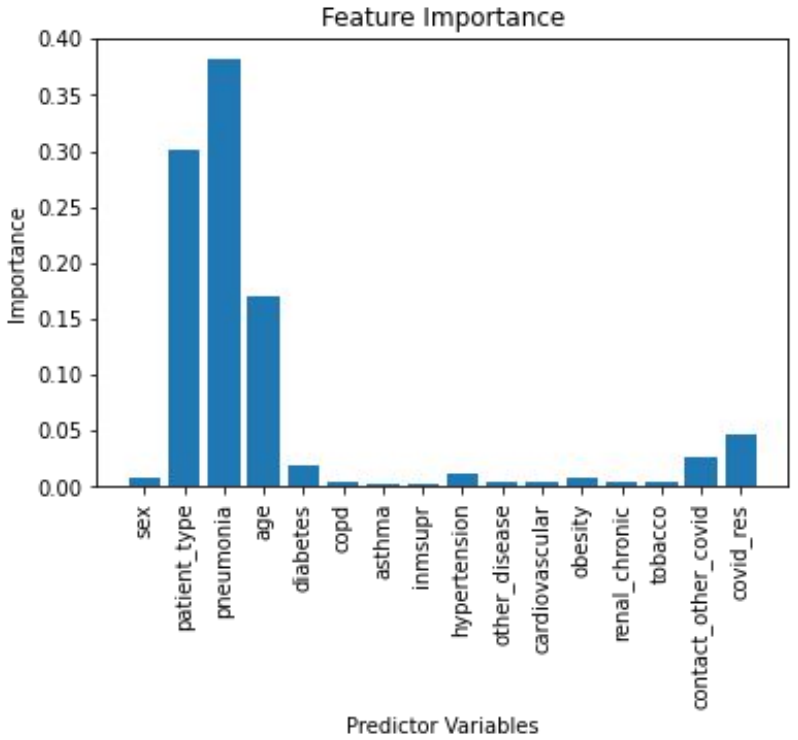
Balanced Dataset	**Logistic Regression**	**Decision Tree**	**Random Forest**
Accuracy	0.9148	0.9513	0.9520
F1 Score	0.9147	0.9513	0.9519

Identifying Our Best Predictors

Using the optimal random forest classifier trained with our balanced data set, we determined patient type (inpatient vs outpatient), pneumonia, and age were the most important predictors.



Same EDA visualization from slide7



Applying our Knowledge

To inform others of the impact pre-existing conditions have on COVID-19 patients, we chose to give the user the ability to explore different conditions and how they impact survival rates with a Command Line Interface (CLI):

```
Hi! Please answer these questions to get a personalized
 probability of survival if hospitalized for COVID-19.

What is your sex? M or F      F

What is your age? Please enter a number:      37

Do you have diabetes? Y or N      Y

Do you have athsma? Y or N      N

.......

Here are your results...

After loading YOUR data into our model, we predict you
have a 72.0% percent chance of survival if hospitalized
 for COVID-19
```

Old Man Hackandcough vs. Young Woman Febreze

Old Man Hackandcough is a 79 year old male previously diagnosed with pneumonia. He was admitted into the hospital for further monitoring and treatment.

Based on our model, Old Man Hackandcough's probability of survival is 65%

Young Woman Febreze is a 24 year old female with no pre-existing health conditions who had an appointment at the hospital for a medical screening.

Based on our model, Young Woman Febreze's probability of survival is 100%.

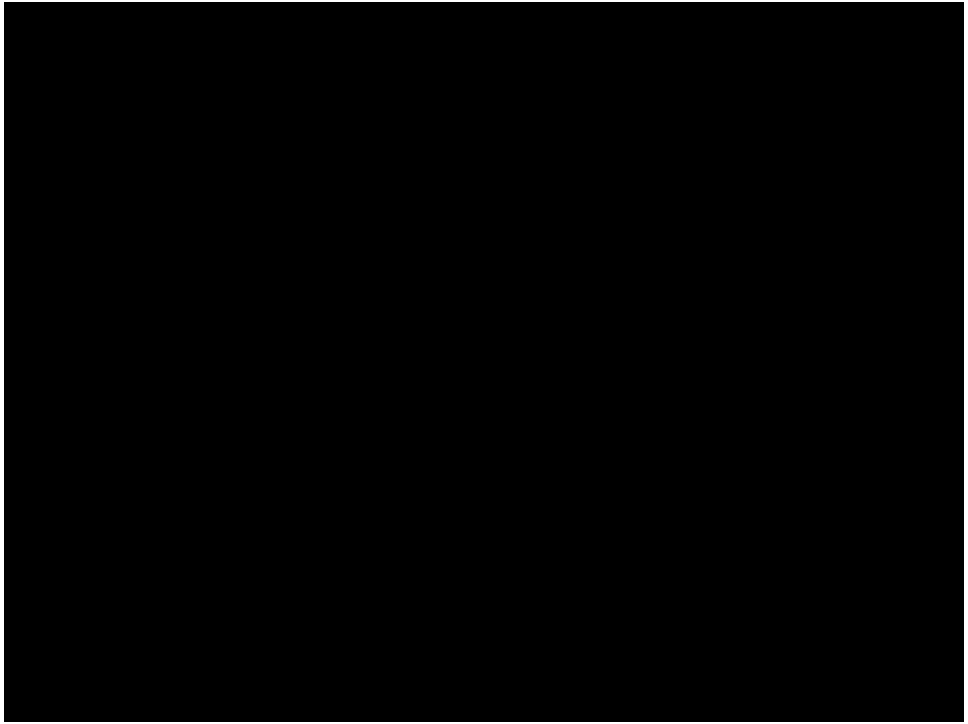
Testing User Input

We used unit testing to ensure the user's data would be properly documented.

```
test_the_age (__main__.MyTest) ... ok
test_the_athsma (__main__.MyTest) ... ok
test_the_cardiovascular (__main__.MyTest) ... ok
test_the_contact (__main__.MyTest) ... ok
test_the_copd (__main__.MyTest) ... ok
test_the_diabetes (__main__.MyTest) ... ok
test_the_hypertension (__main__.MyTest) ... ok
test_the_obesity (__main__.MyTest) ... ok
test_the_otherdisease (__main__.MyTest) ... ok
test_the_pneumonia (__main__.MyTest) ... ok
test_the_renalchronic (__main__.MyTest) ... ok
test_the_sex (__main__.MyTest) ... ok
test_the_tobacco (__main__.MyTest) ... ok

----------------------------------------------------------------------
Ran 13 tests in 0.026s

OK
<unittest.main.TestProgram at 0x7fcf3bd05050>
```



Summary

The aim of this project was to gain an understanding of how pre-existing conditions impact patients with COVID-19 in hospitals.

We set out to accomplish this by comparing mortality rates, exploring visualizations, and making predictions using a data set obtained from the Mexican government.

We learned that many pre-existing conditions impact COVID-19 patients.

Based on our model, we identified pneumonia, patient type, and age to be the most important when predicting the probability of survival.

We then conveyed our understanding effectively by allowing the user to input different pre-existing conditions to obtain the probability of survival.