THE EFFECT OF HEALTH CONDITIONS ON COVID-19 DEATH RATES

Digital Futures

Katie Jones

Agenda

INTRODUCTION

BACKGROUND ON THE DATA

ACHIEVING A MODEL

OPTIMUM MODEL

MOST AT RISK

LIMITATIONS

CONCLUSIONS

Introduction

Covid-19 is still a relatively new topic that has caused around 6.66 million deaths worldwide.

Using anonymized patient-related information from a Mexican dataset, health conditions that put you most at risk of dying from coronavirus will be predicted.

What I will be talking about today:

- Insight into the dataset and visualisations
- Explanation of the Prediction Model
- Conditions predicted that put patients at risk

Covid-19 in Mexico

Total cases – 7.16 million

Total deaths – 331,000

The first official cases were confirmed in February 2020

The true impact of the pandemic in Mexico is massively underestimated.

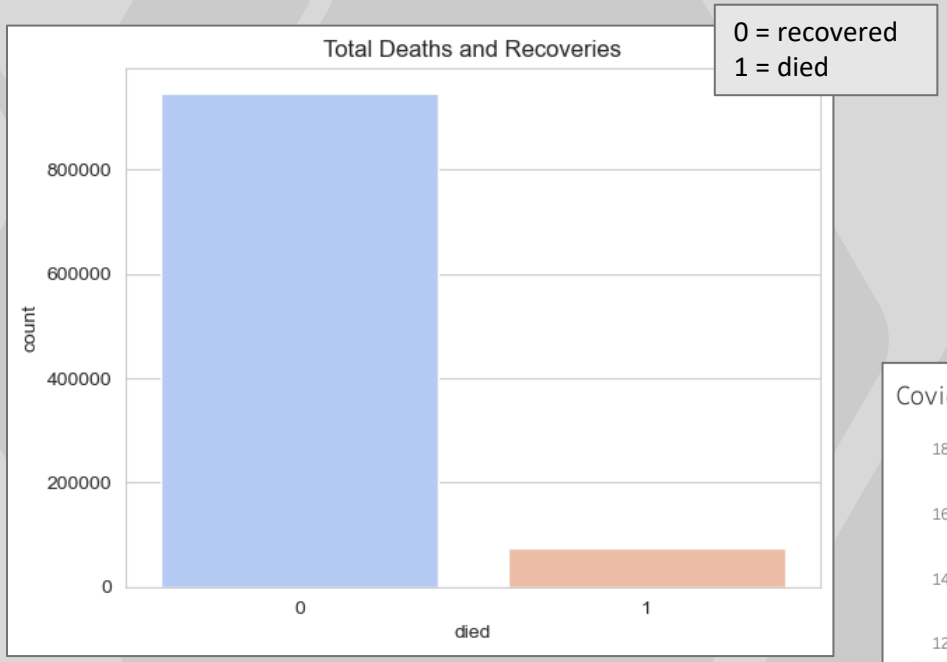
Mexico was considered to have a low testing rate in comparison to other countries – especially at the beginning of the pandemic.

The Dataset

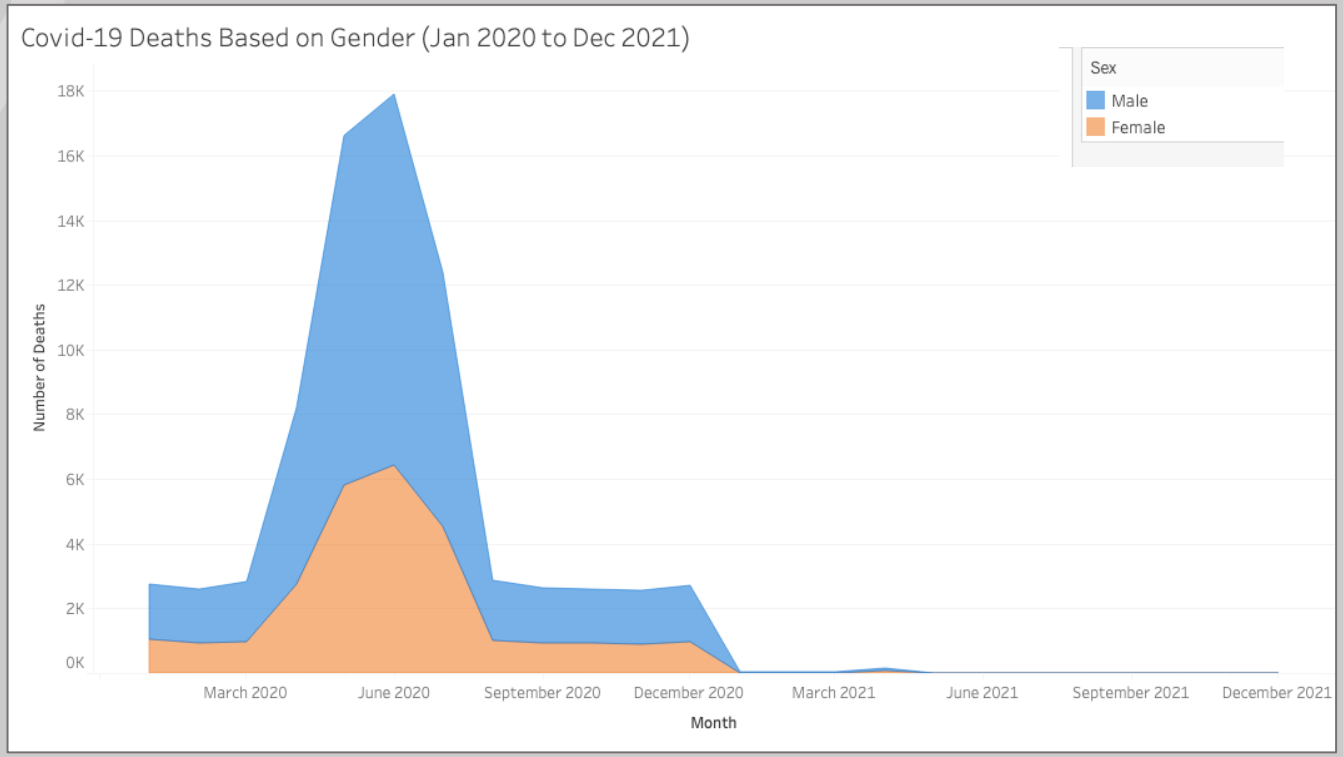
- Provided by the Mexican Government
- Over a million patients from 2020 and 2021
- Has information on pre-existing conditions
- Determines whether the patient died or recovered
- Need to work out a model to determine at risk patients

Condition	% Died
Chronic Renal Disease	30.19
COPD	26.7
Diabetes	22.61
Cardiovascular	21.35
Hypertension	19.7
Immunosuppressed	18.48
Other Disease	16.21
Obesity	10.82
Tobacco Users	7.82
Asthma	4.69
Pregnancy	1.09

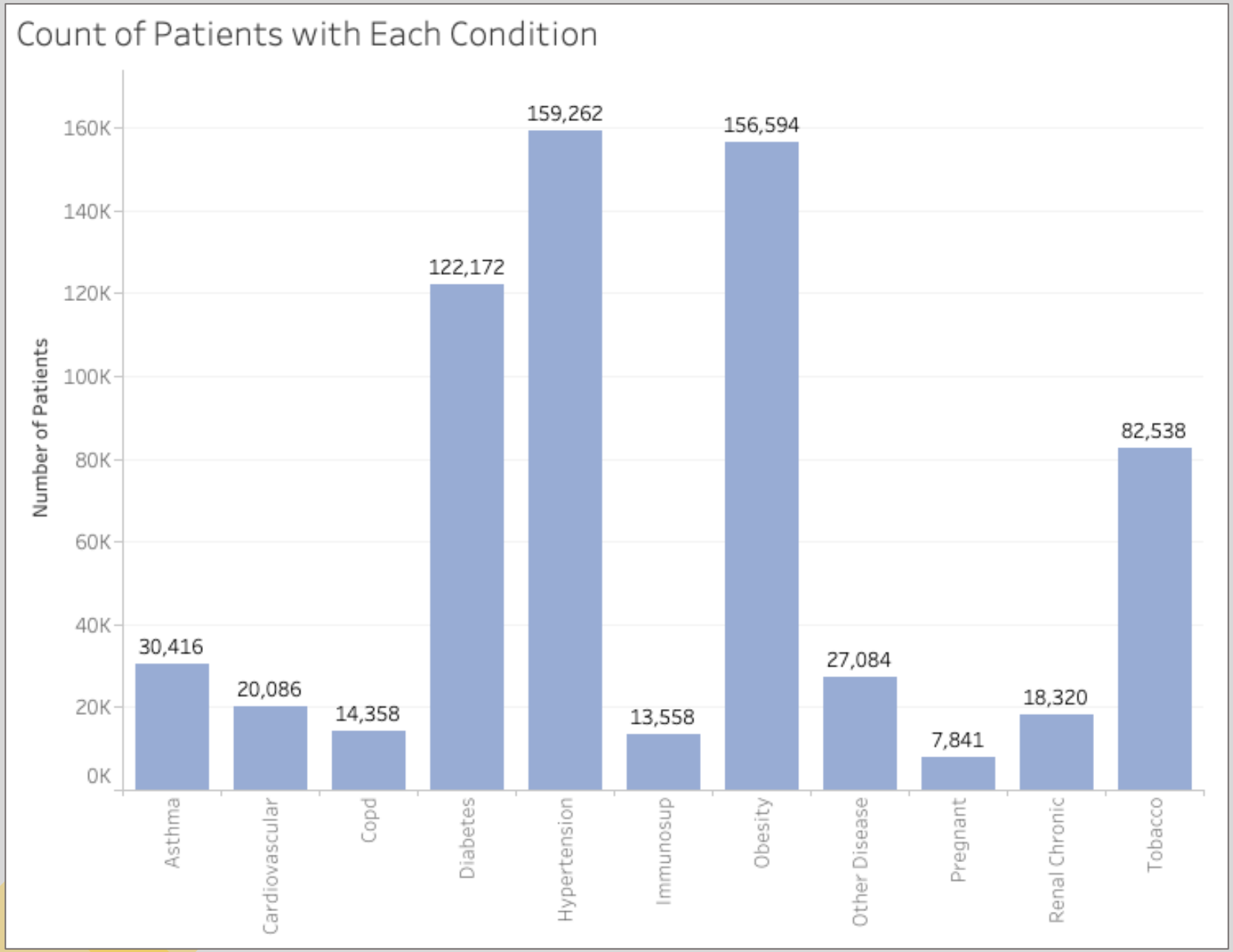
Deaths



- A much larger proportion of patients that died were males.
- The majority of deaths occurred in June 2020.



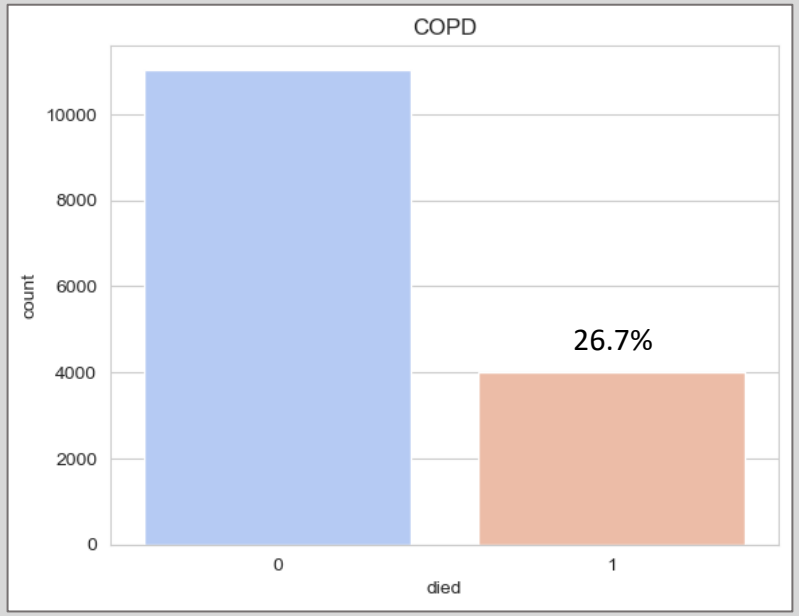
- Out of the million patients in the dataset, 7.31% of them died and 92.69% recovered.

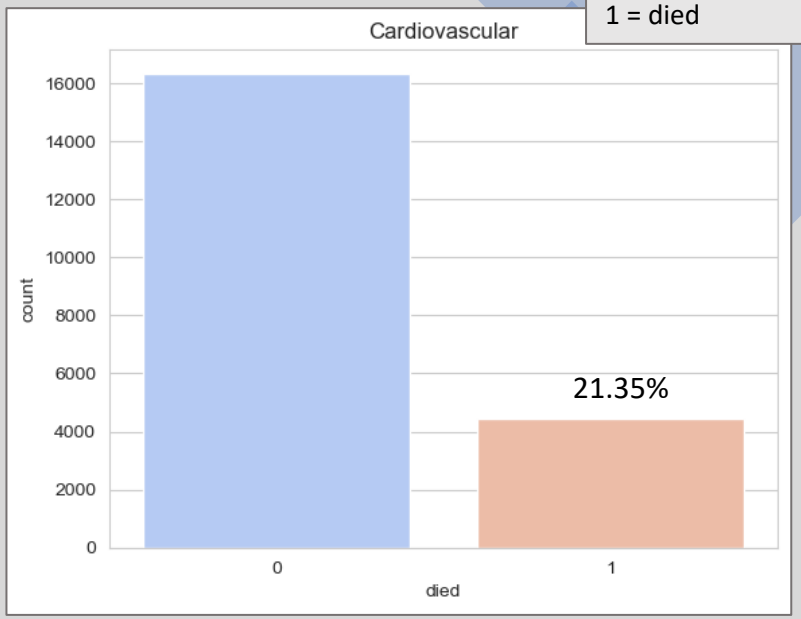

Count of Patients with Each Condition
Number of Patients
160K
140K
120K
100K
80K
60K
40K
20K
0K
30,416
20,086
14,358
122,172
159,262
13,558
156,594
27,084
7,841
18,320
82,538
Asthma
Cardiovascular
Copd
Diabetes
Hypertension
Immunosup
Obesity
Other Disease
Pregnant
Renal Chronic
Tobacco

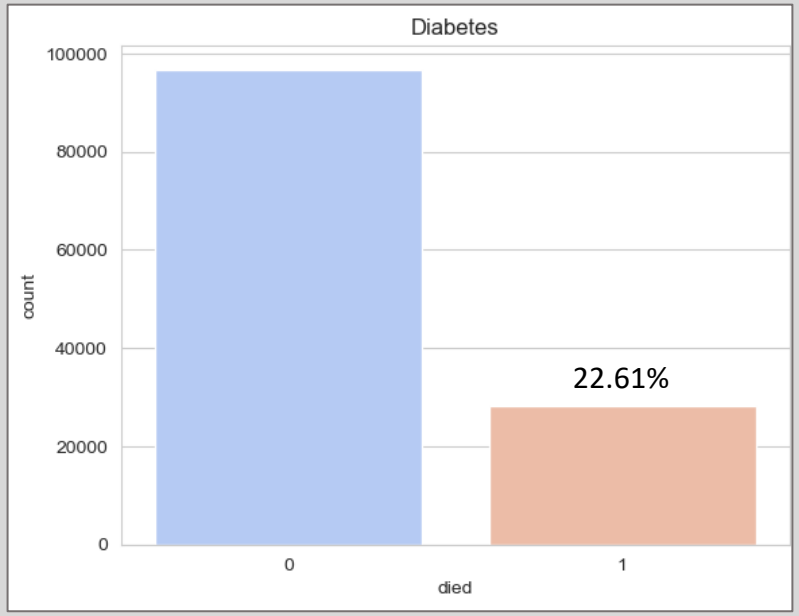
Conditions with High Deaths

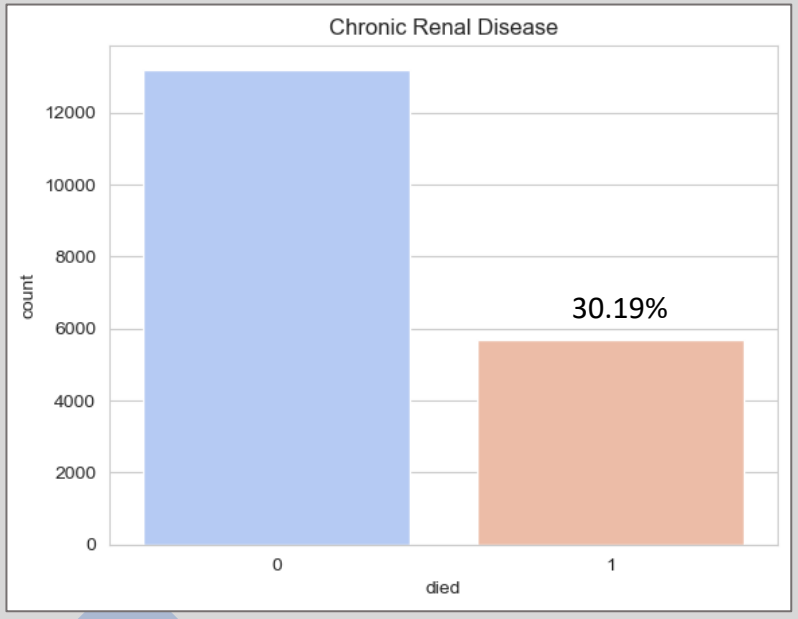
- These show the breakdown of patients that died and recovered who had:
  - COPD
  - Cardiovascular disease
  - Diabetes
  - Chronic Renal Failure
- These conditions had the highest proportions of deaths with over 20% of patients with each condition dying.

0 = recovered
1 = died









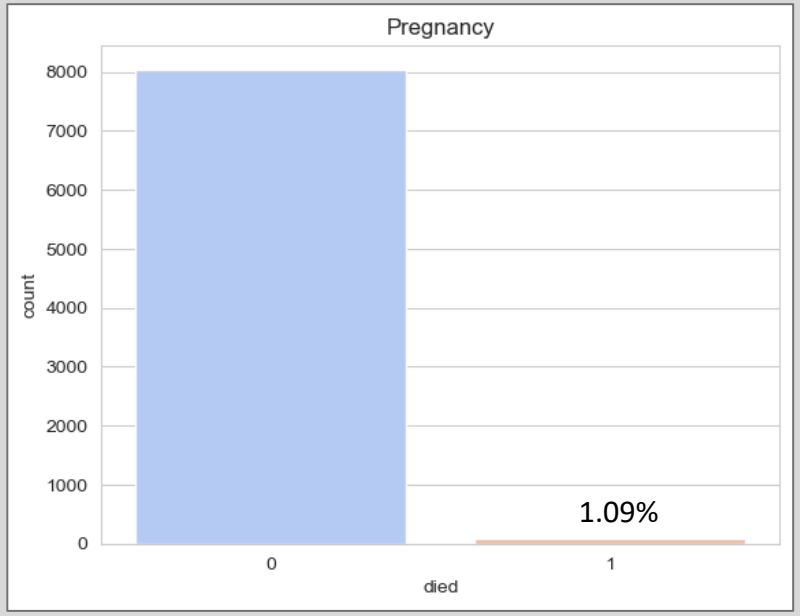
Conditions with Low Deaths

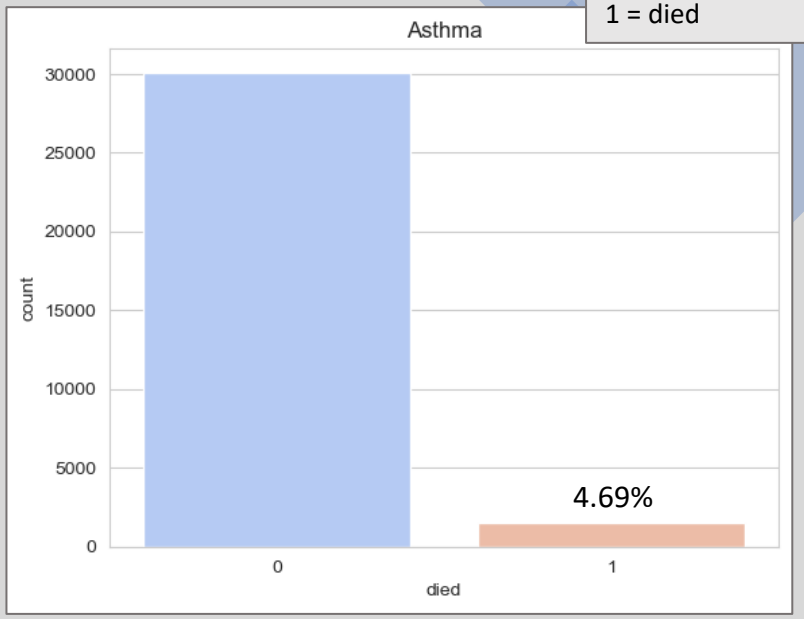
- These show the breakdown of patients that died and recovered who:
  - Were pregnant
  - Had asthma
  - Were tobacco users
  - Were obese

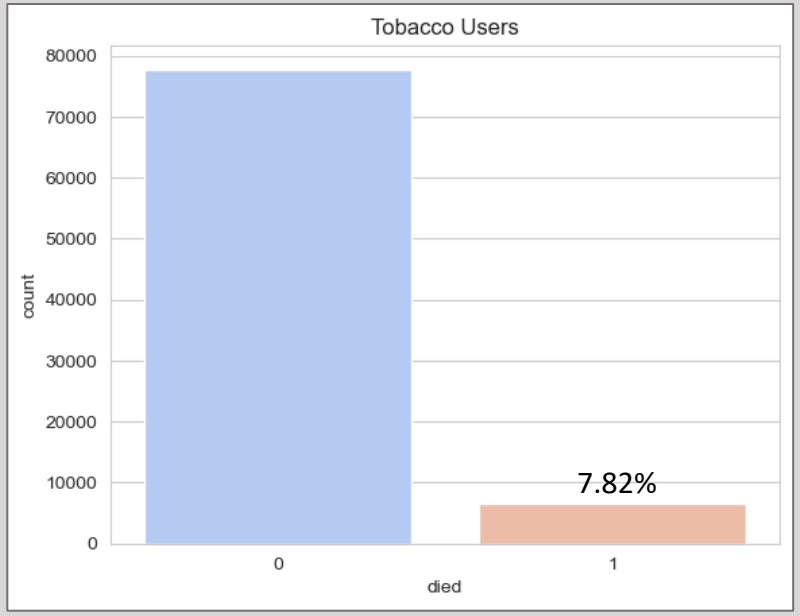
- These conditions had the lowest proportions of deaths.

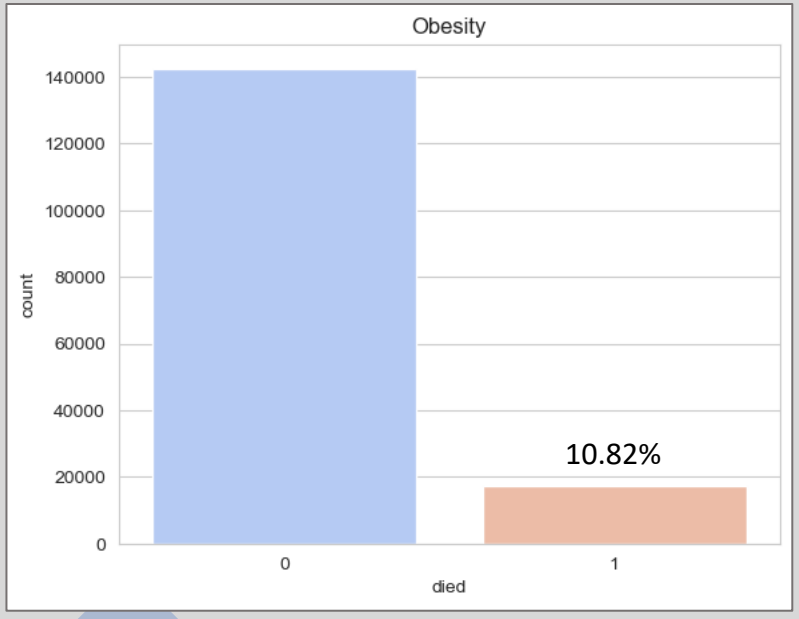
- The number of pregnant people in the overall dataset was very low.

0 = recovered
1 = died

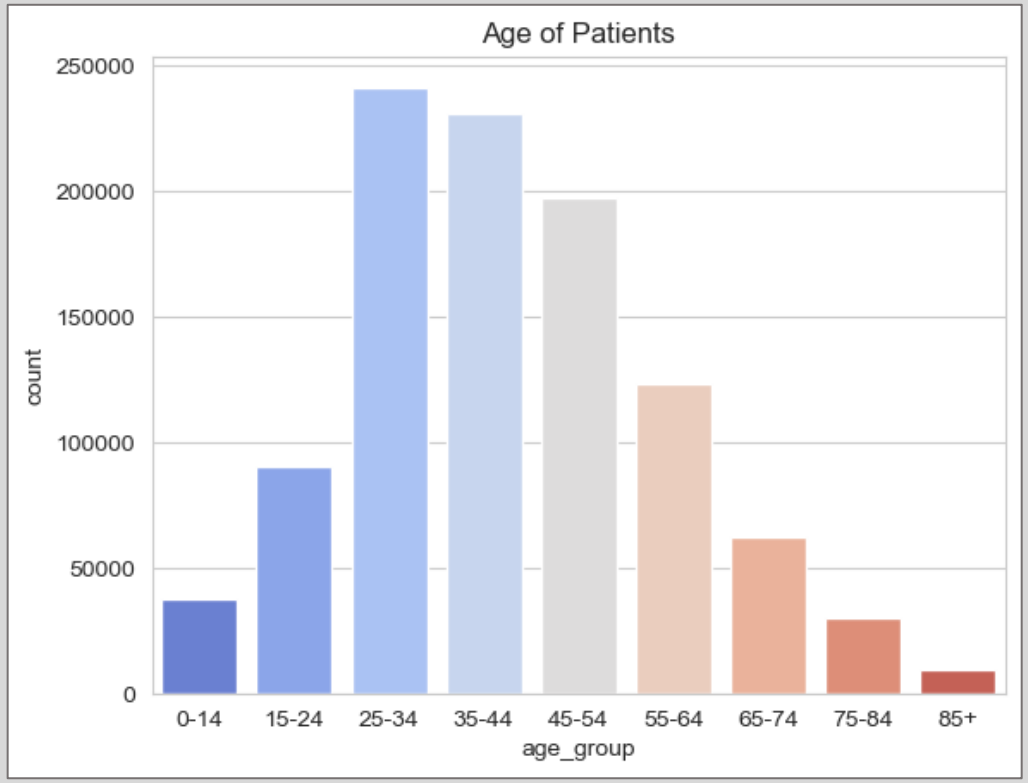


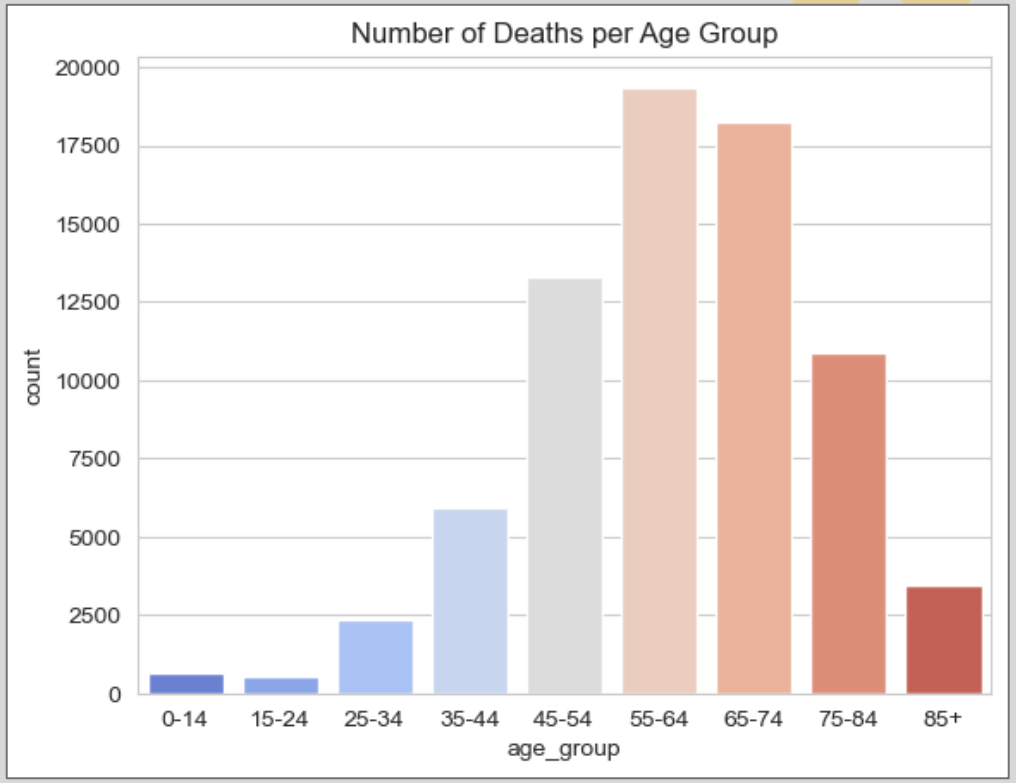






Age Groups



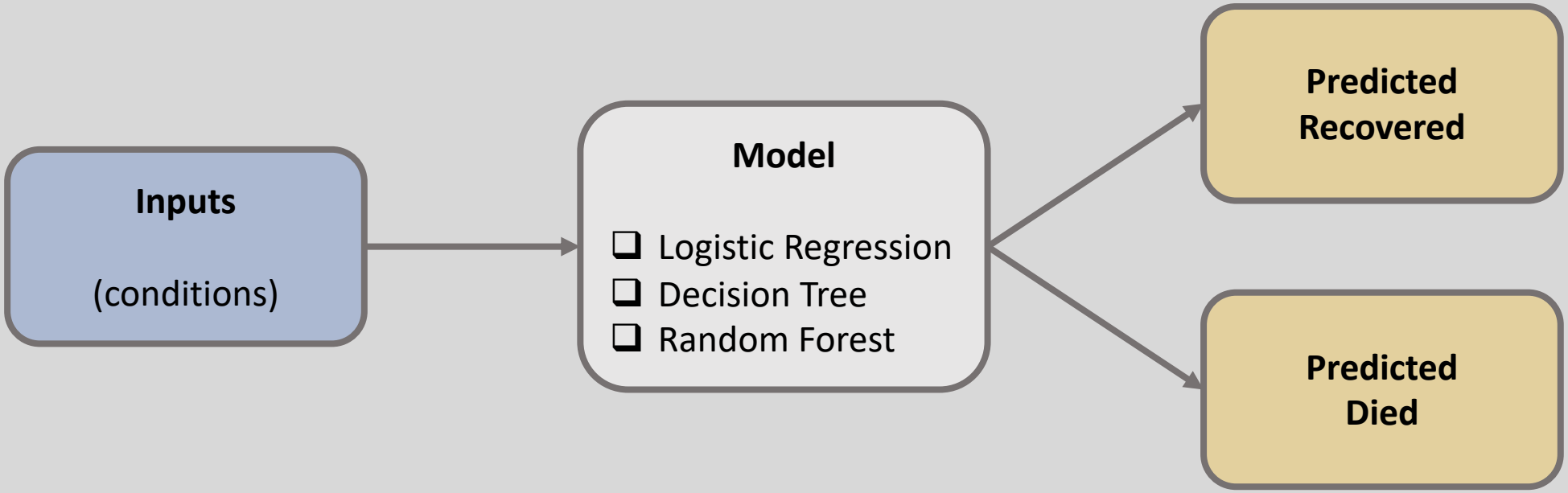


The largest age group in the dataset is 25-34

The age group with the most deaths is 55-64 followed closely by 65-74

75-85 and 85+ also have large numbers of deaths despite their group size

Achieving a Model



- I used three models with different variations of data:
  - Sample of 1% of the data
  - Undersampled data – a balance of ‘died’ and ‘recovered’
  - Different feature columns
- I also merged the three models together to get the mode ‘y_pred’
- Once I had a large spread of models, I assessed which was the best version to use.

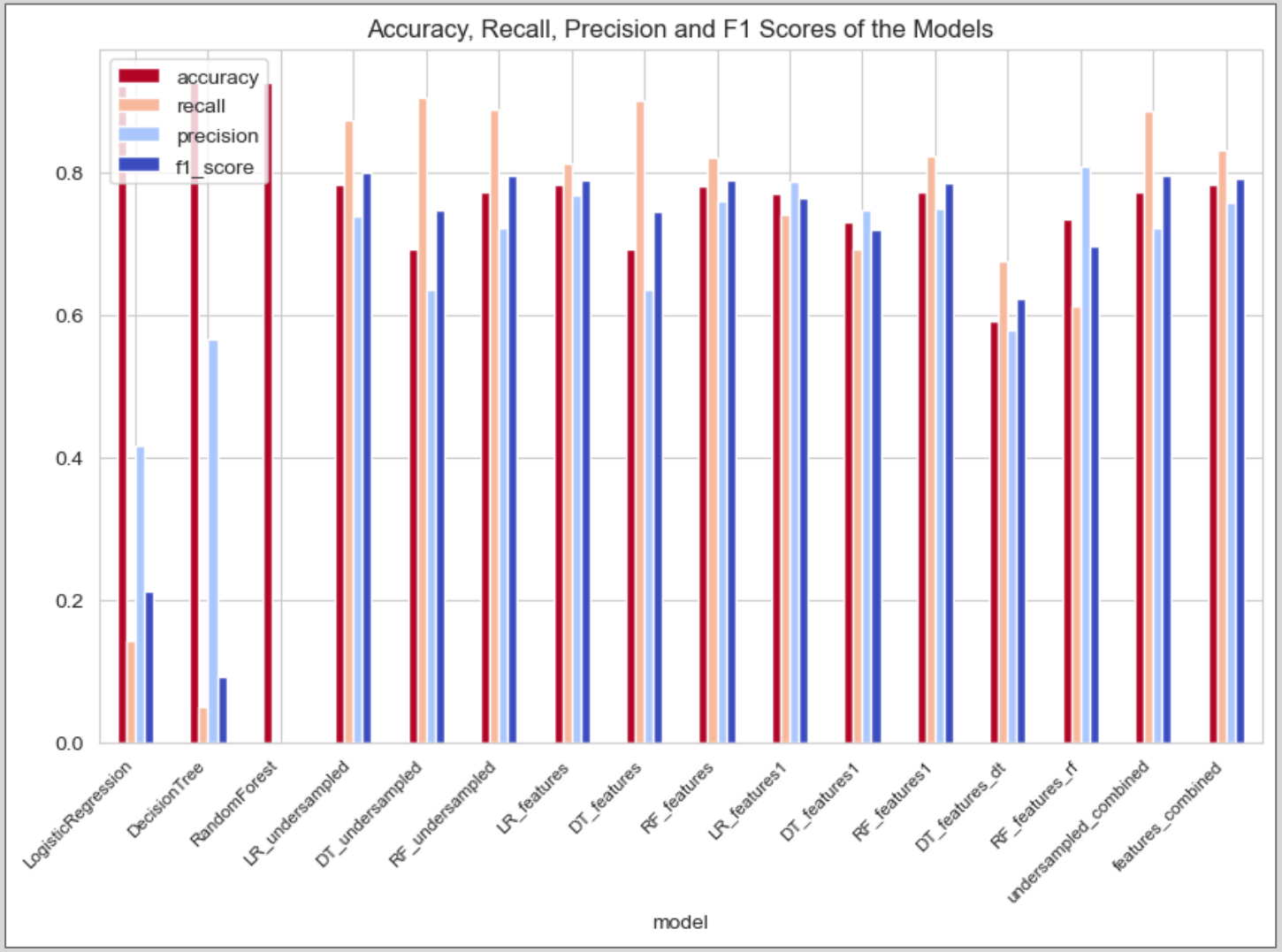
Choosing the Optimum Model

- Ideally, we want all 4 scores to be as high as possible
- A high recall is most important
- A high accuracy is also important as it tells us what percentage we predicted correctly

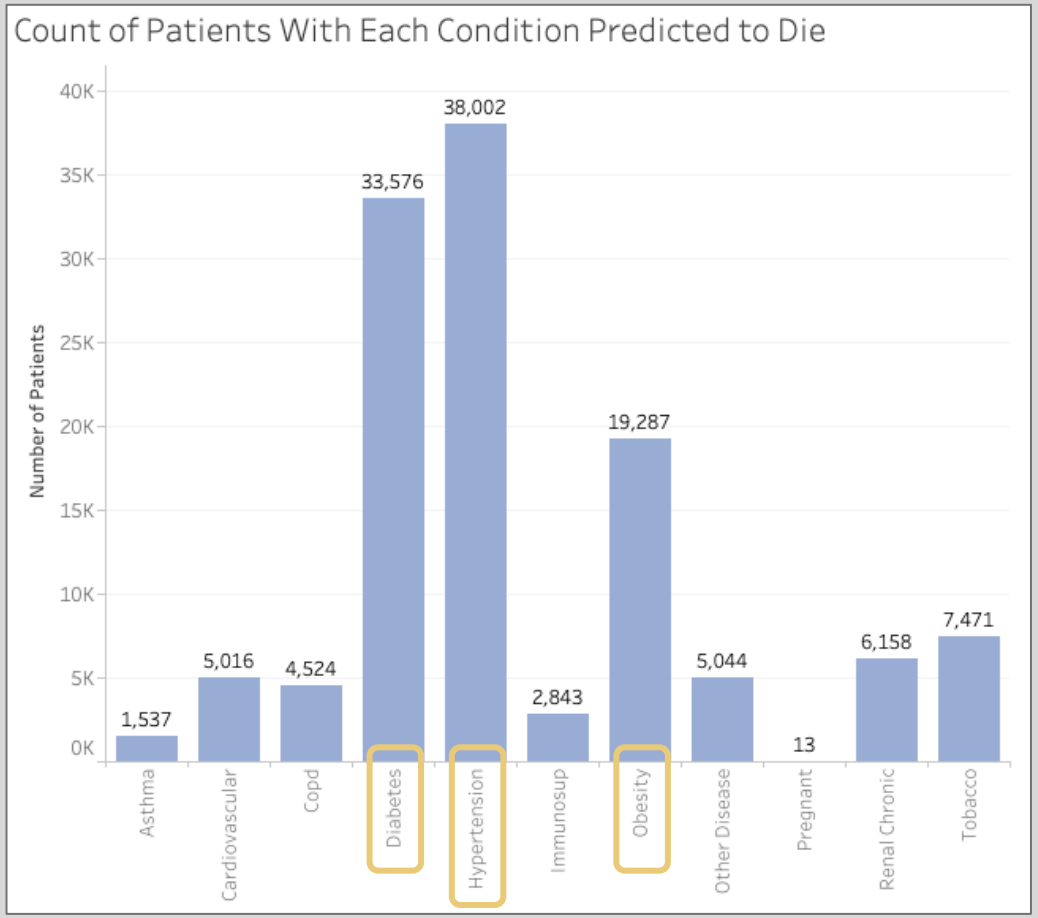
Logistic Regression on the undersampled dataset using all of the columns was the best model to use.

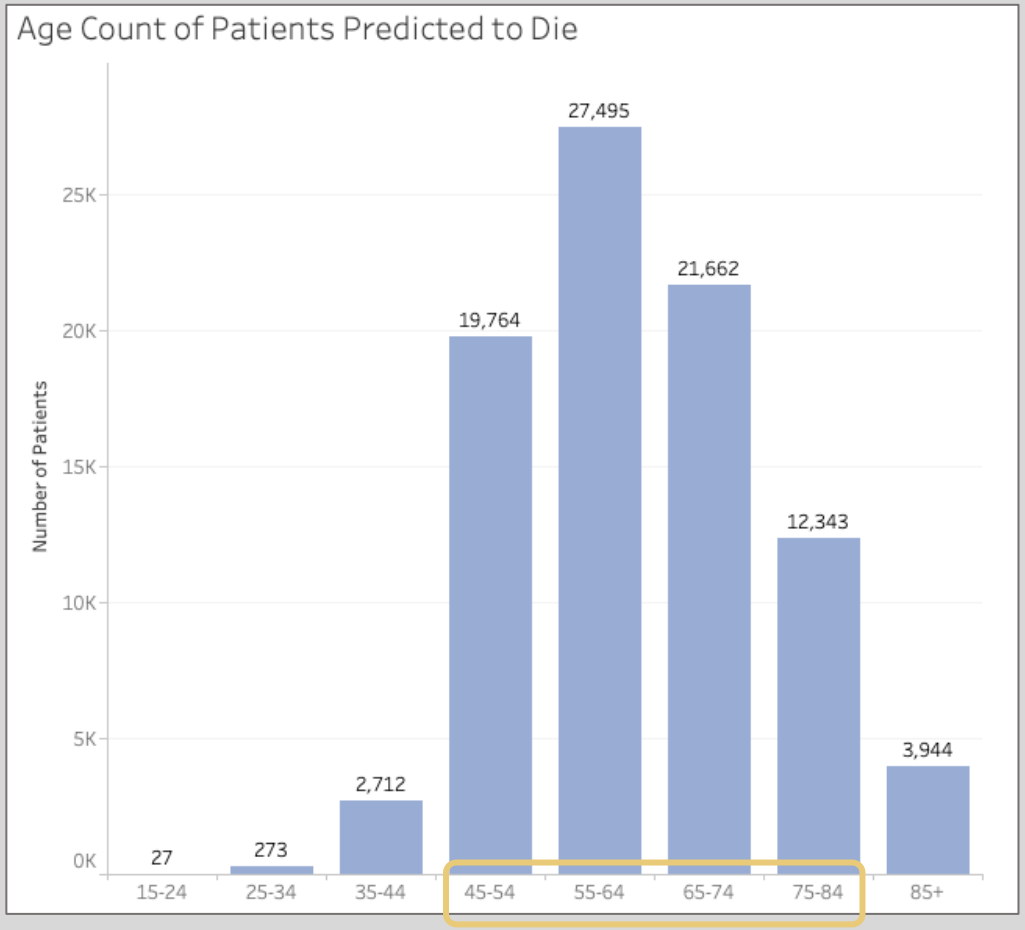
```
The confusion matrix for your predictions is:
[[10385  4598]
 [ 1900 12980]]

The accuracy of your model is: 0.7824063222047349
The recall of your model is: 0.8723118279569892
The precision of your model is: 0.7384230287859824
The F1-score of your model is: 0.799802822108571
```



What makes a patient at risk based on my model?



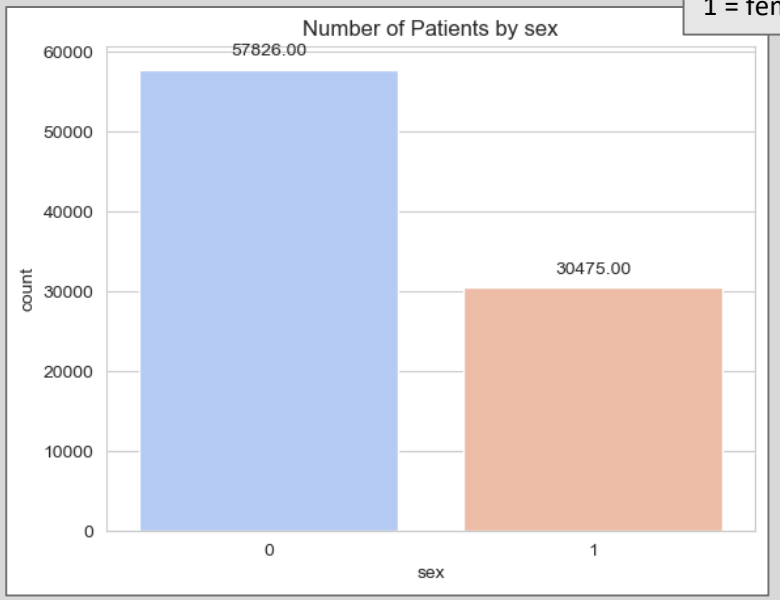


What makes a patient at risk based on my model?

The greater the coefficient (both positive and negative) the higher the importance

Condition	Coefficient
Renal chronic	1.29
Immunosuppressants	0.88
Sex	-0.81

0 = male
1 = female



65% of the patients predicted to die from Covid-19 were males

Limitations

- It is a Mexican dataset and we know Mexico had one of the lowest test rates so we cannot be sure how accurate the data is.

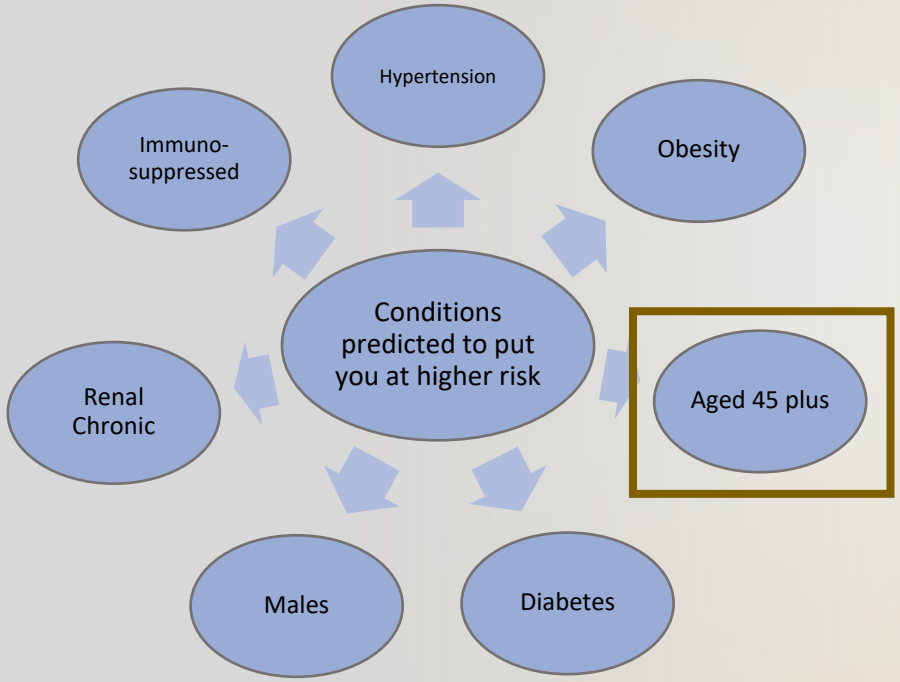
- The imbalance between ‘died’ and ‘recovered’ made it challenging to provide a reproducible model
  - Using the random undersampler helped with this but if I had had more time, I would have used other methods such as oversampling (SMOTE) or an algorithm approach and compared the results.

- I predicted 9467 to recover but they died
  - If I had more time I would have tried other models to see if I could bring this number down.

- Covid-19 is still very much a new concept.

Summary

- The best model for this dataset was Logistic Regression



- These predictions can be used to determine treatment plans for patients

Digital Futures

THANK YOU

ANY QUESTIONS?