COVID-19 Death Prediction

Henry McTernan & Maris Ryan

COVID-19 Background

Global

- Started in early 2020
- COVID-19 is a contagious, respiratory disease in which people with immunocompromised or other illnesses may be more likely to be at serious risk for a worse infection
- Almost 7 million deaths from COVID-19 around the world
    - 772,000,000 cases
- 13,595,583,125 vaccine doses

Mexico

- 334,786 deaths from COVID-19
    - 7,693,120 cases
- Fifth most COVID-19 related deaths
- 222,921,381 vaccine doses
- Seventh highest mortality rate
    - ~4.3%

COVID-19 Dataset

- Kaggle
- Released by the Mexican Government in 2022
- The dataset used 1,048,576 observations and has 21 predictors

Predictors

- Date_Died
  - 9999-99-99 did not die, otherwise died
- USMER
  - Medical levels (1, 2, or 3)
- Medical_Unit
  - type of institution of the National Health System that provided the care.
- Sex
  - 1 female, 2 male
- Patient_Type
  - 1 for returned home, 2 for hospitalized
- Intubed
  - If patient needed a ventilator
- Pneumonia
  - whether the patient already has air sacs inflammation or not.
- Age
- Pregnant
  - whether the patient is pregnant or not.
- Diabetes
  - whether the patient has diabetes or not.
- COPD
  - Chronic obstructive pulmonary disease or not.
- Classification
  - Values 1-3 mean that the patient was diagnosed with covid in different degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive.
- Asthma
  - whether the patient has asthma or not.
- Inmsupr
  - whether the patient is immunosuppressed or not.
- Hypertension
  - whether the patient has hypertension or not. (when the pressure in your blood vessels is too high (140/90 mmHg or higher))
- Cardiovascular
  - whether the patient has heart or blood vessels related disease.
- Renal chronic
  - whether the patient has chronic renal disease or not.
- Other disease
  - whether the patient has other disease or not.
- Obesity
  - whether the patient is obese or not.
- Tobacco
  - whether the patient is a tobacco user.
- ICU
  - Indicates whether the patient had been admitted to an Intensive Care Unit.

Research Questions

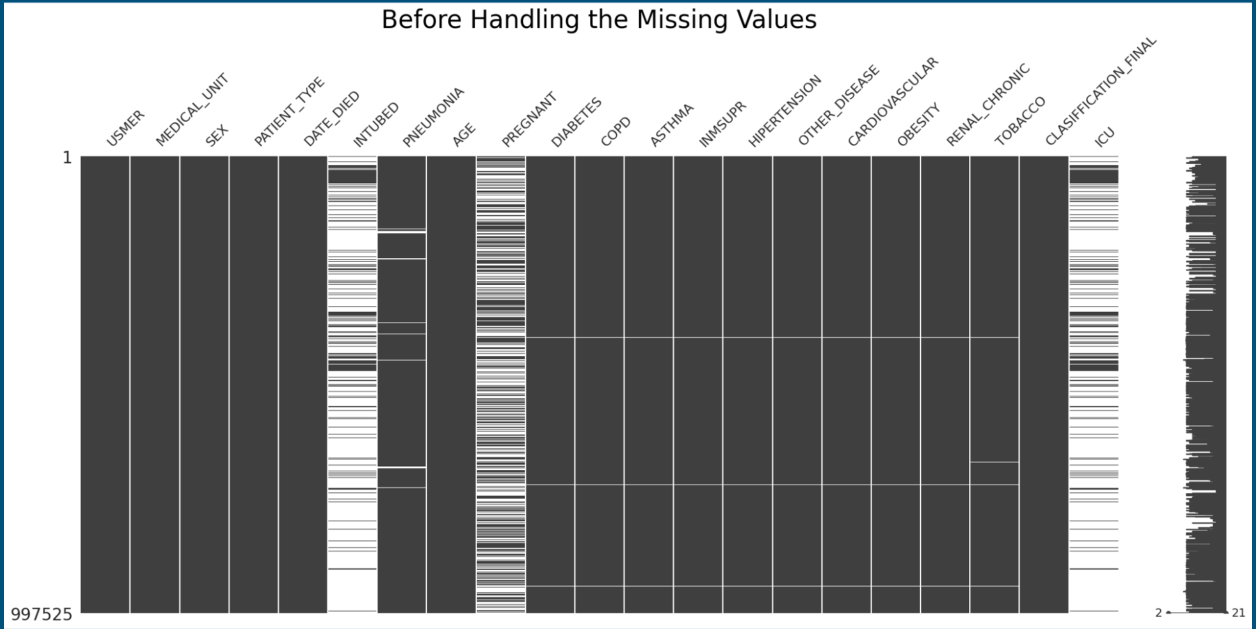
1. Which machine learning model gives the best prediction of death in patients with COVID-19?
    a. How well can the model's predictions be explained and interpreted, especially in the context of healthcare decision-making?
2. Which predictors are the best at predicting a patient’s death from COVID-19?
    a. How does the inclusion/exclusion of specific predictors affect the algorithm's performance?

Preliminary Studies

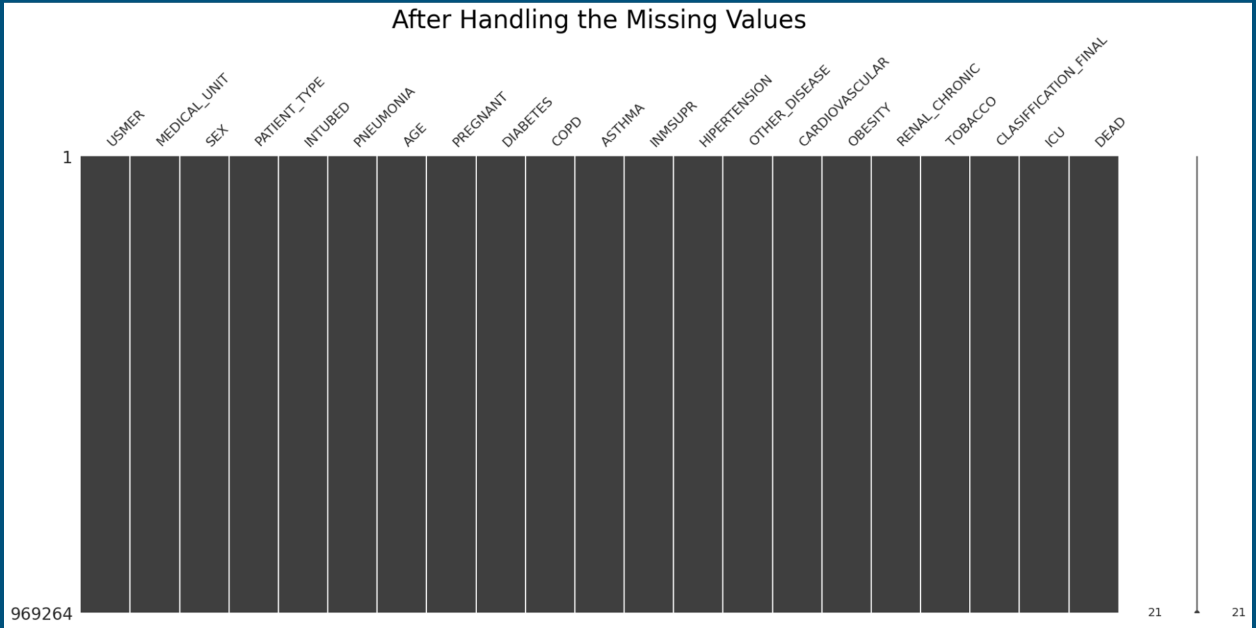
Issues with our Data

- Many missing values
- Imbalance in the dataset
- Many variables have a low correlation with our target variable

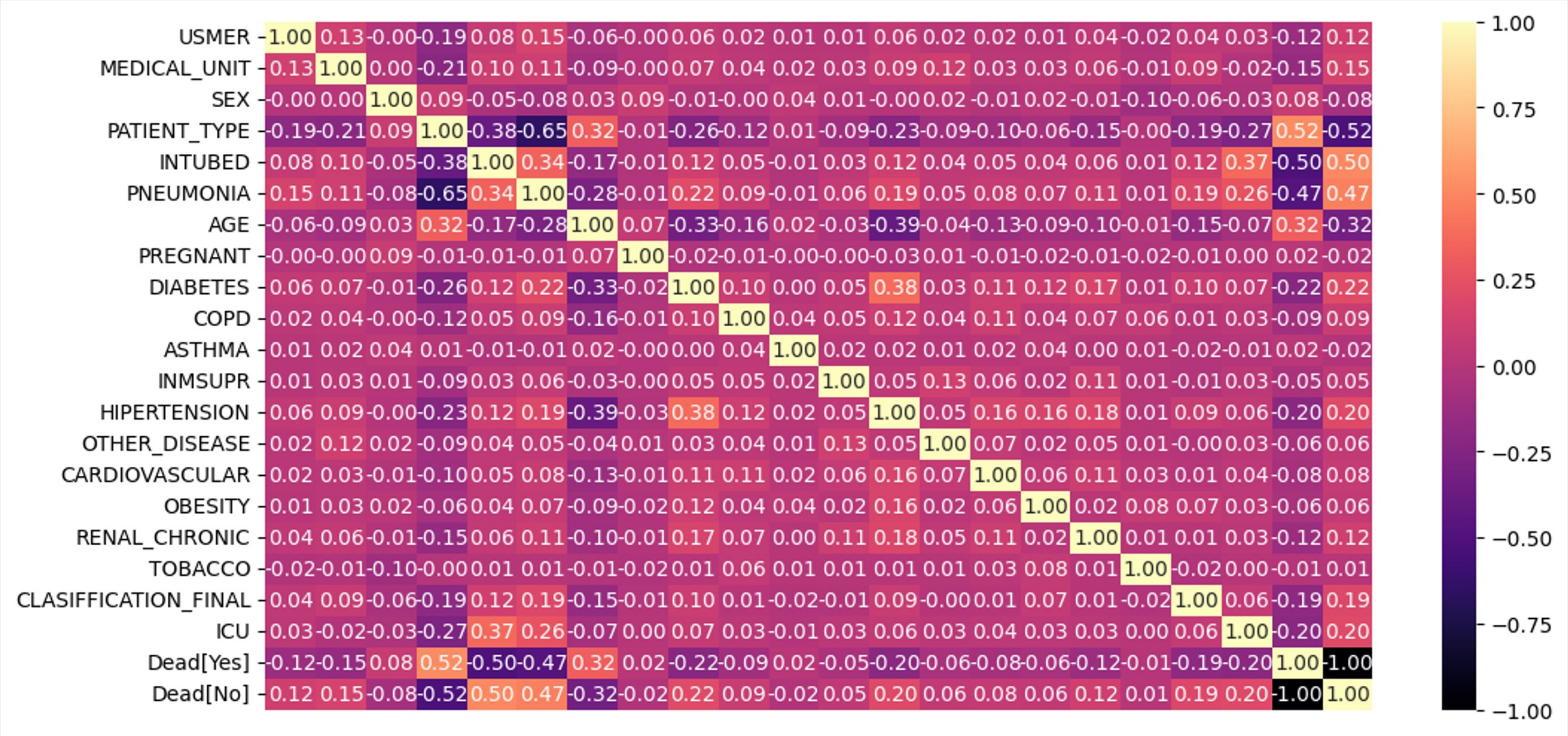
Handling Missing Values



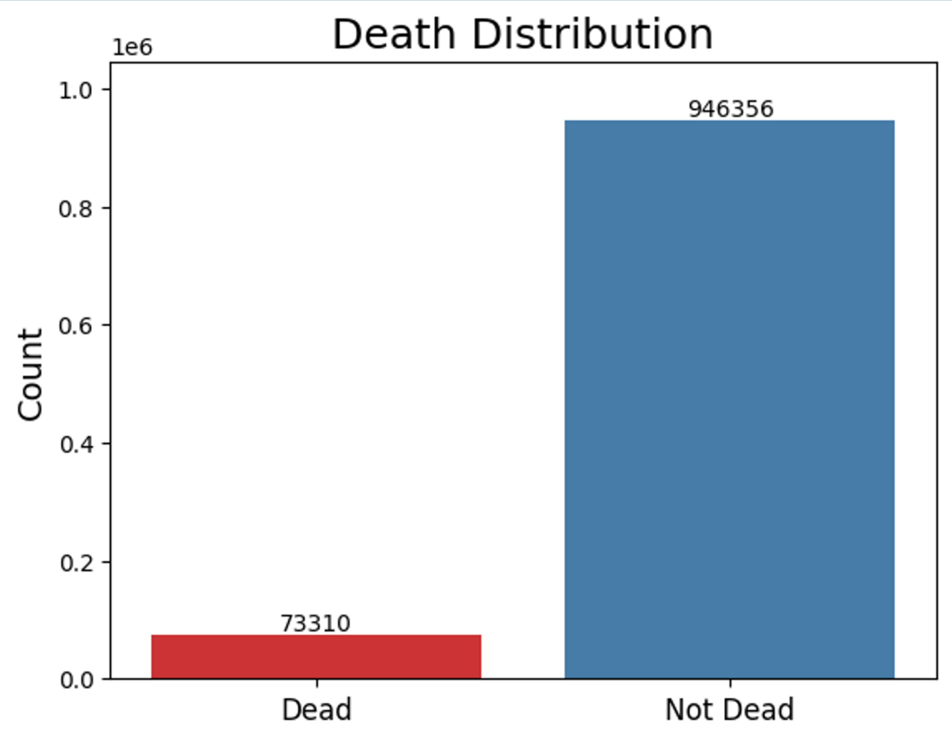
Handling Missing Values



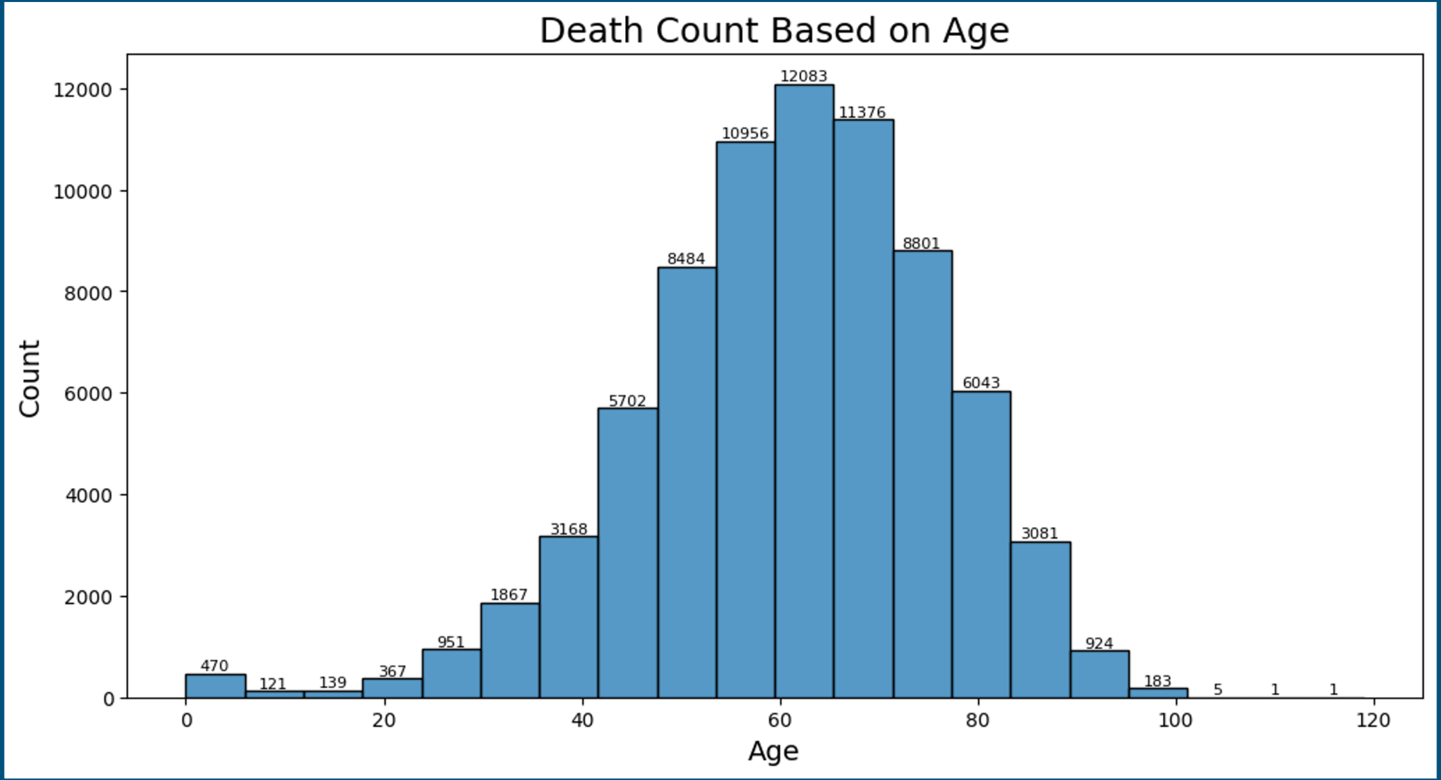
Correlation of Variables



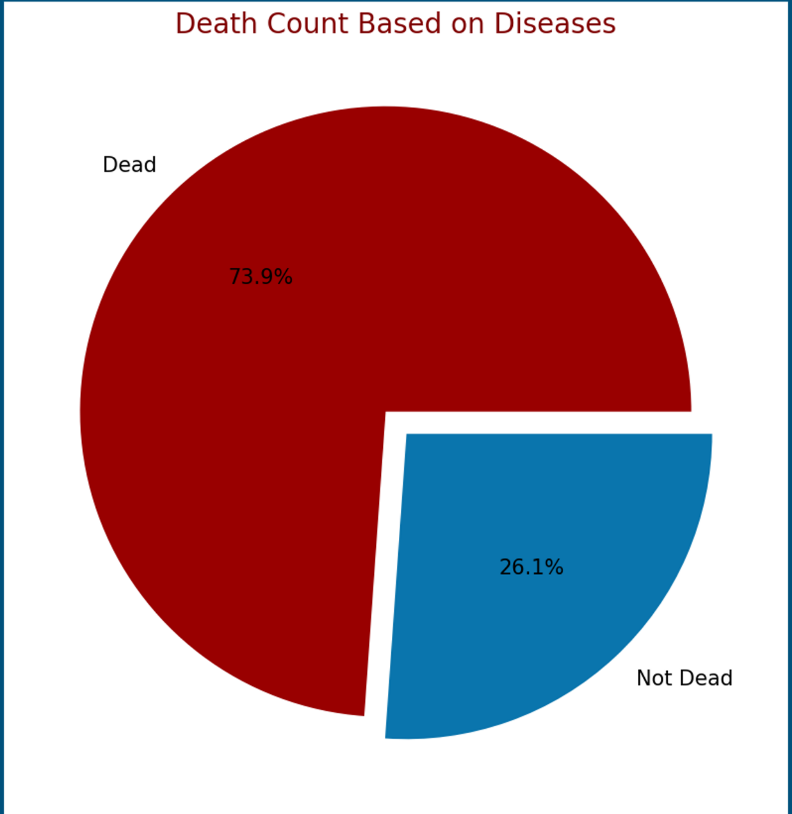
How many people have died?



Which age group has the most deaths?



Does having a disease affect percentage of death?



Which diseases were considered:

- Pneumonia
- Diabetes
- Chronic obstructive pulmonary disease
- Asthma
- Immunosuppressed
- Cardiovascular related disease
- Chronic renal disease
- Obesity
- Other

Statistical Analysis Methods & Results

Solving Imbalance in a Dataset

- Loading More Data
- Changing The Performance Metrics
- Resampling (Undersampling or Oversampling)
- Changing The Algorithm
- Penalized Models etc.

Resampling

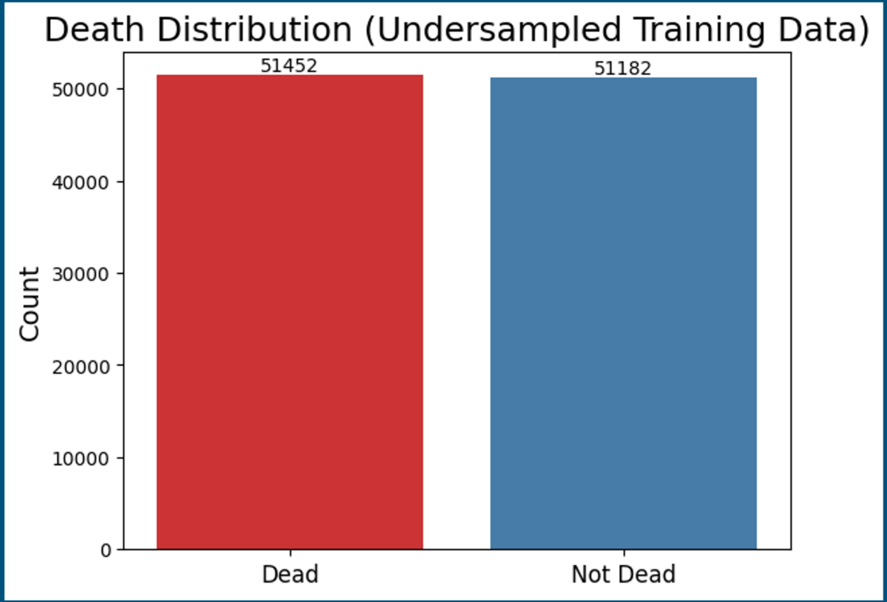
Undersampling:

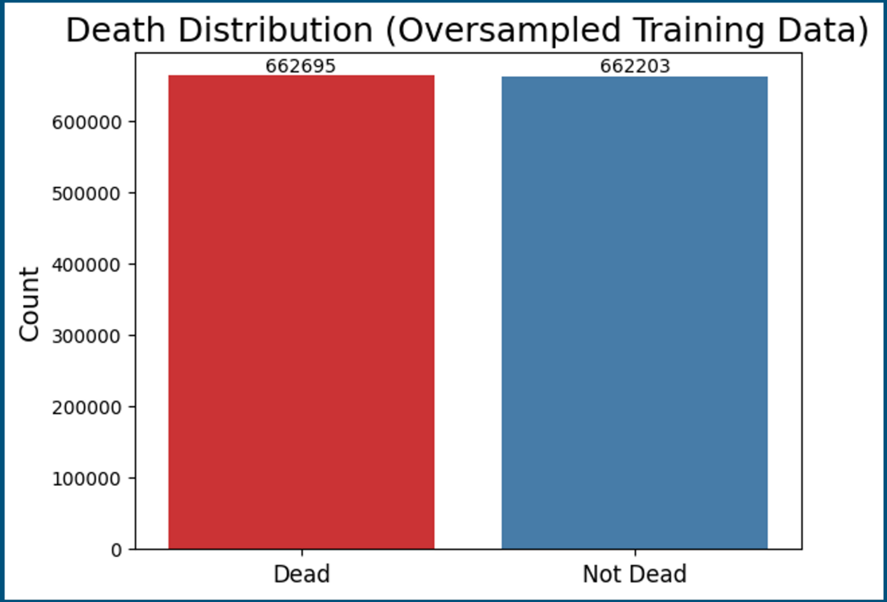
- Modify the distribution of a variable in your dataset by artificially decreasing the number of observations that take on a particular value or range of values for that variable
- Deleting samples from the majority class (‘Not Dead’)
- Pros
    - Does not introduce repeated or redundant information
- Cons
    - Reduces the size of your dataset
    - Loses potentially valuable information

Oversampling:

- Modify the distribution of a variable in your dataset by artificially increasing the number of observations that take on a particular value or range of values for that variable
- Duplicating samples from the minority class (‘Dead’)
- Pros
    - Do not lose any information
- Cons
    - Increases the chance of overfitting
    - Increases the learning time of the training data

After Resampling

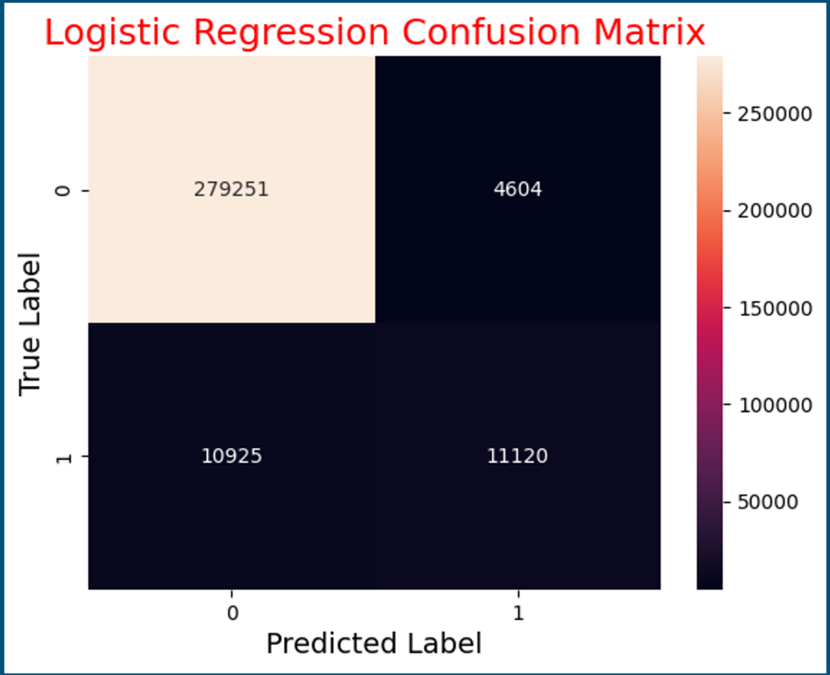




Logistic Regression

- Aims to find the best fitting model to describe the relationship between the dependent variable and independent variable(s)
- One of the most simple machine learning models
  - Easy to interpret and very efficient to train
- Works more efficiently when you remove variables that have no or little relation to the output variable

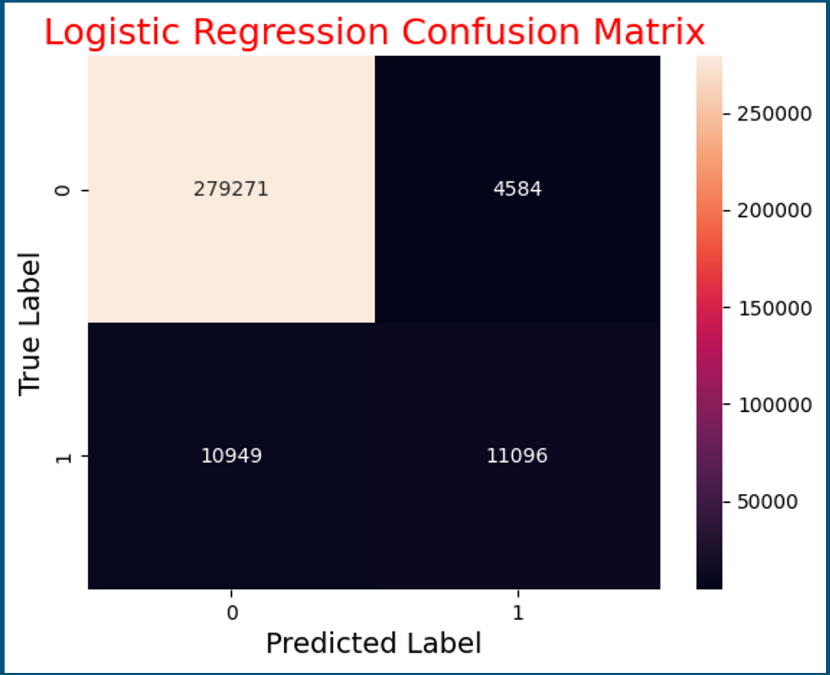
Removing missing values:



Accuracy = 0.9492350441320693
Precision = 0.7071991859577715
Recall = 0.5044227716035382

FPR = 0.29280081404222846
FNR = 0.03764956440229378

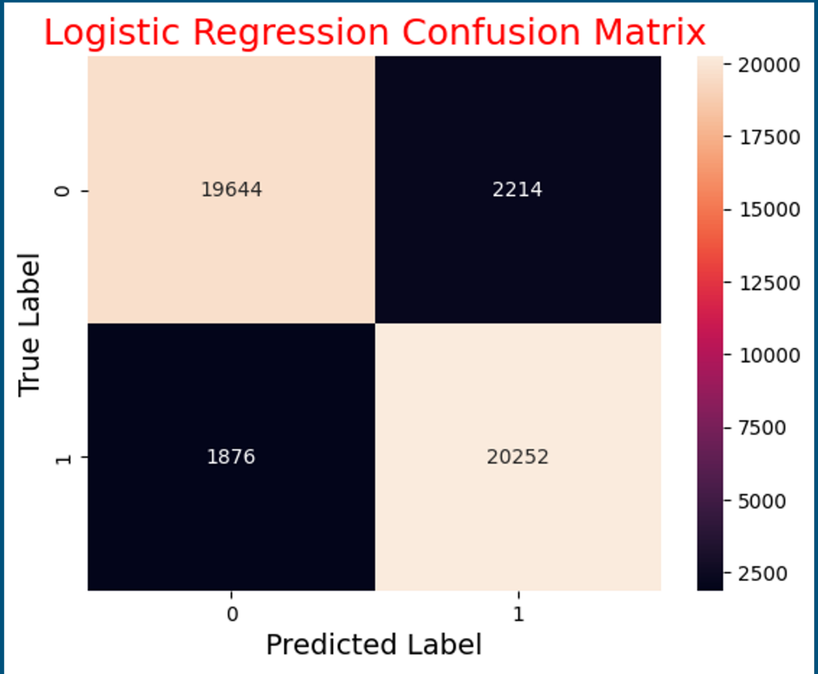
Removing irrelevant variables:



Accuracy = 0.9492219679633868
Precision = 0.7076530612244898
Recall = 0.5033340893626673

FPR = 0.2923469387755102
FNR = 0.0377265522706912

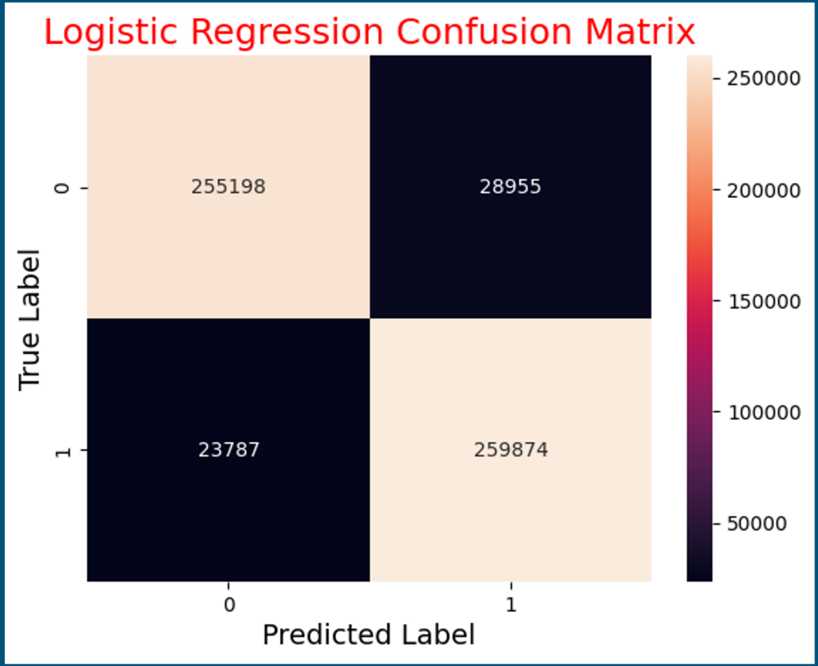
Undersampling:



Accuracy = 0.9070158686854909
Precision = 0.9014510816344699
Recall= 0.9152205350686913

FPR = 0.09854891836553013
FNR = 0.08717472118959108

Oversampling:



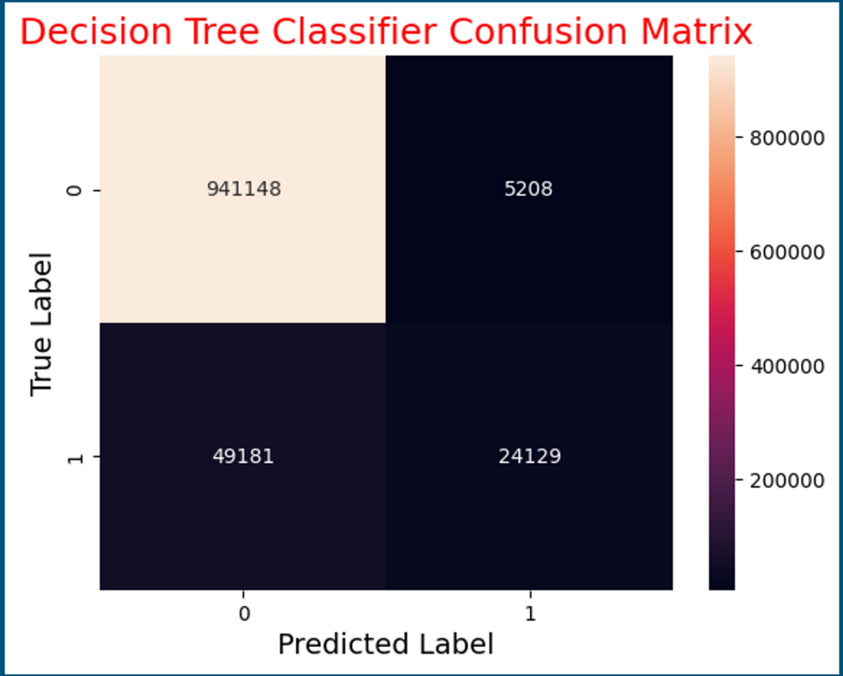
Accuracy = 0.9071139492862099
Precision = 0.899750371326979
Recall = 0.916142860668192

FPR = 0.10024962867302106
FNR = 0.08526264852949084

Decision Tree Classification

- Goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features
- Supervised learning algorithm wherein the data points are continuously split according to certain parameters and/or the problem that the algorithm is trying to solve
- Uses a data structure called a tree to predict the outcome of a particular problem
- Non-linear – can capture complex relationships and interactions between features

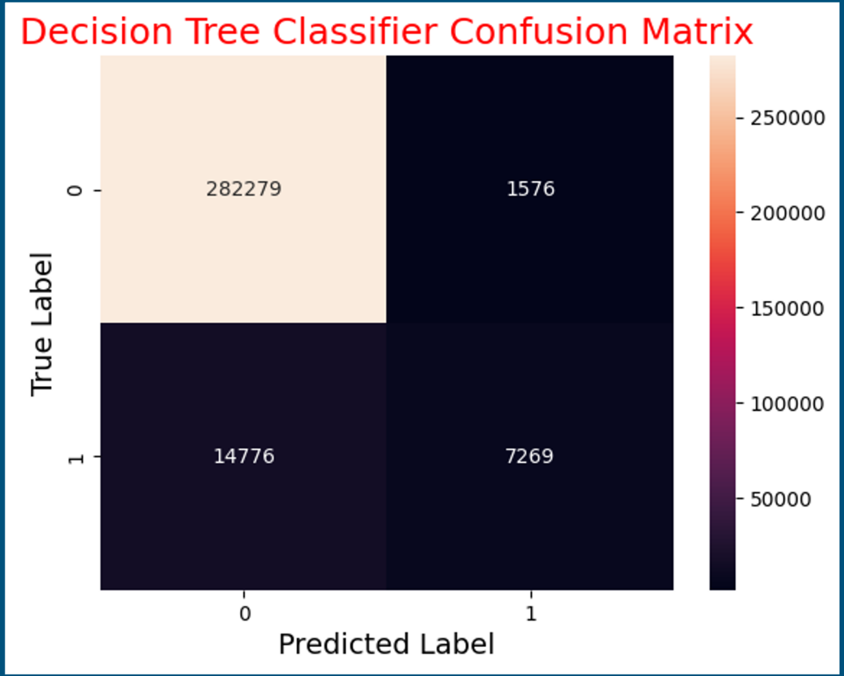
From cleaning missing values:



Accuracy = 0.9466599847401012
Precision = 0.65
Recall = 0.52

FPR = 0.17752326413743735
FNR = 0.04966127418262012

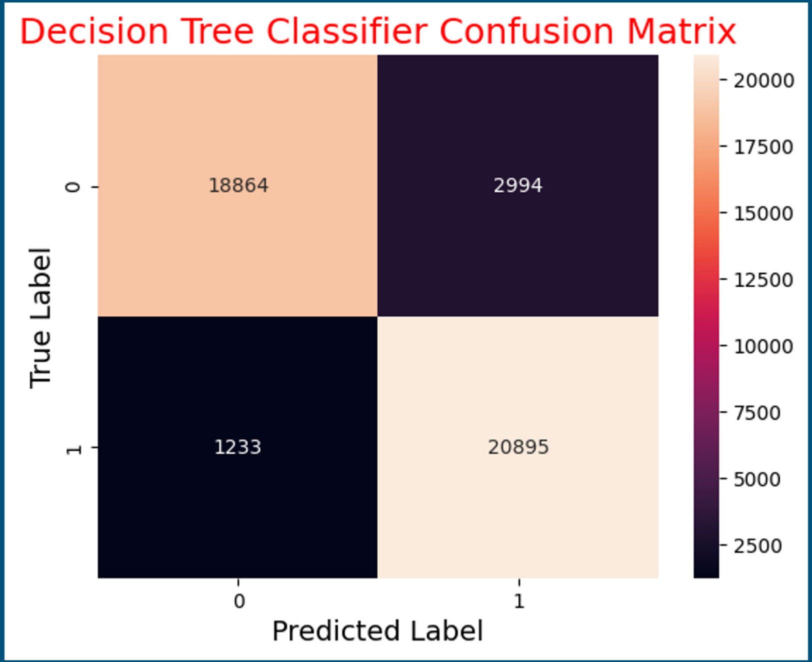
Removing irrelevant variables:



Accuracy = 0.9465446224256293
Precision = 0.64
Recall = 0.53

FPR = 0.17817976257772752
FNR = 0.3431889443734758

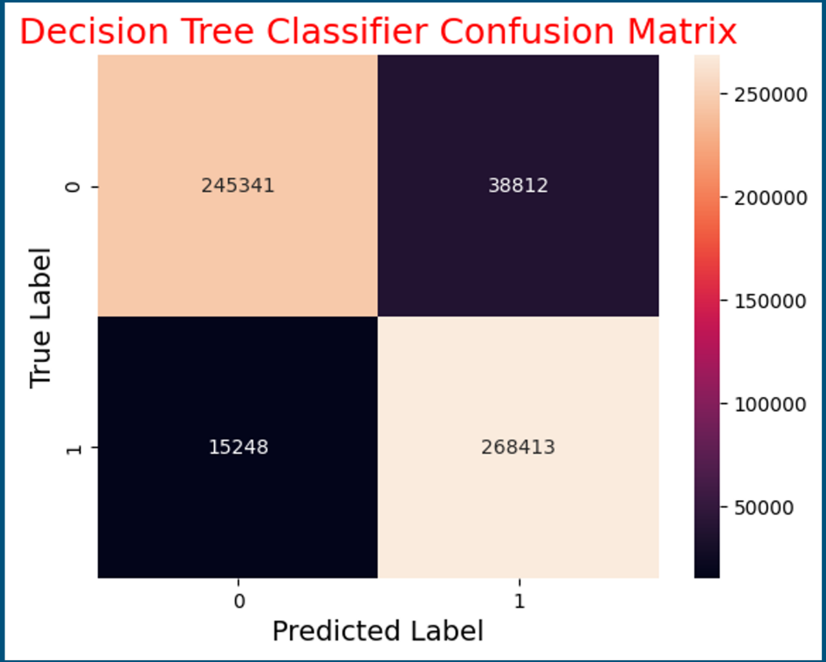
Undersampling:



Accuracy = 0.9039012413040513
Precision = 0.64
Recall = 0.53

FPR = 0.1253296496295366
FNR = 0.06135244066278549

Oversampling:



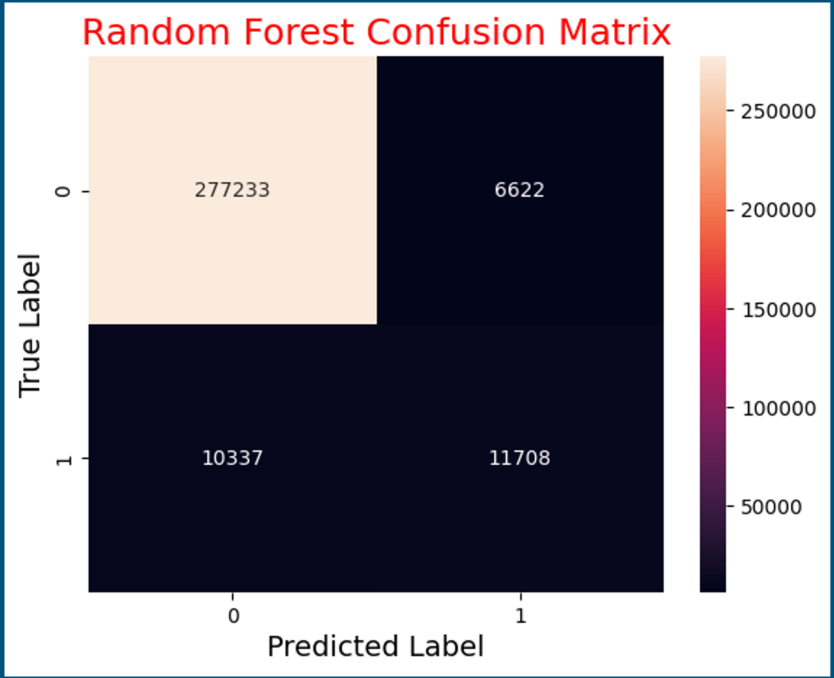
Accuracy = 0.9047927666454155
Precision = 0.89
Recall = 0.92

FPR = 0.1263308650012206
FNR = 0.05851359804136015

Random Forest Classification

- Since our DTC's aren't very precise and have either a high FPR or FNR, we decided Random Forest would be another model to test because it can often improve performance and is less prone to overfitting
- Enables any classifiers with weak correlations to create a strong classifier
- Good at handling large datasets
- Superior method for working with missing data because missing values are substituted by the variable appearing the most in a particular node

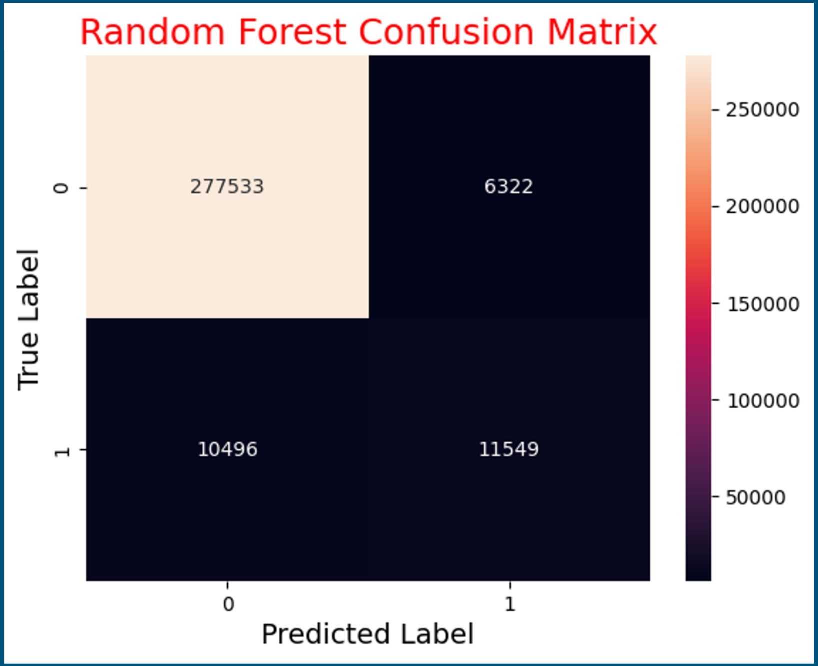
Removing missing values:



Accuracy = 0.9445603138280484
Precision = 0.64
Recall = 0.53

FPR = 0.3612568466993997
FNR = 0.03594603053169663

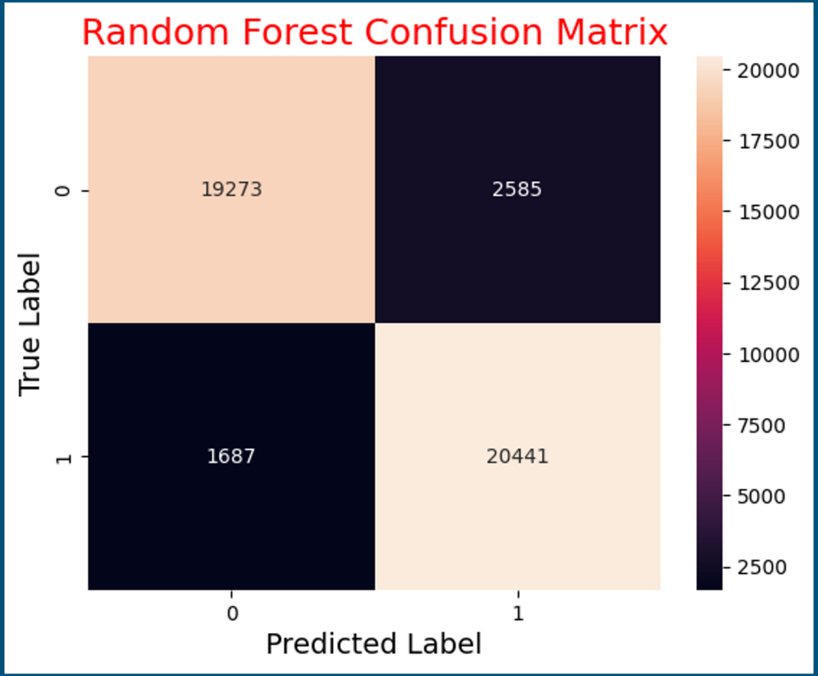
Removing irrelevant variables:



Accuracy = 0.9450212487741092
Precision = 0.65
Recall = 0.52

FPR = 0.3537574841922668
FNR = 0.03644077506084457

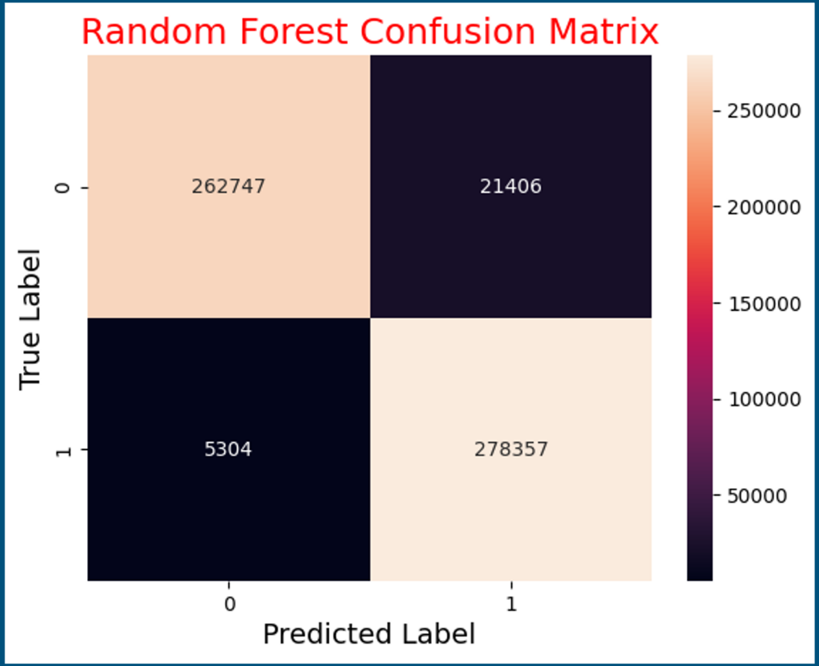
Undersampling:



Accuracy = 0.9028781885145274
Precision = 0.89
Recall = 0.92

FPR = 0.11226439676886997
FNR = 0.08048664122137404

Oversampling:



Accuracy = 0.9529599481520357
Precision = 0.93
Recall = 0.98

FPR = 0.07140974703348979
FNR = 0.019787279286404454

Discussion

- Best model
    - Random forest after oversampling
        - 95.3% accurate
        - FPR of ~7.1%
        - FNR of ~2.0%
    - Address overfitting
        - Training accuracy = 0.964548969052712
        - Test accuracy = 0.9529599481520357
- How our prediction can help
    - Upon improvement, our model can enable early identification of high-risk individuals, allowing healthcare professionals to intervene promptly and initiate appropriate treatments
    - Can guide further research into the specific factors influencing COVID-19 mortality

Future Research

- How can we optimize our results
    - Try other ways to balance our data
    - Try more machine learning models to get our accuracy closer to 100%
- Applying aspects of our model to help predict the risk of death for other health crises
    - Infectious disease outbreaks: Flu, Ebola, Zika, different COVID-19 variants, or future pandemics

Thank you!
Questions?

Sources

https://coronavirus.jhu.edu/data/mortality

https://towardsdatascience.com/the-perfect-recipe-for-classification-using-logistic-regression-f8648e267592

https://www.kdnuggets.com/2022/04/logistic-regression-classification.html

https://towardsdatascience.com/an-exhaustive-guide-to-classification-using-decision-trees-8d472e77223f

https://corporatefinanceinstitute.com/resources/data-science/random-forest/

https://crunchingthedata.com/oversampling-vs-undersampling/