DIAGNOSTIC EFFICIENCY WITH COVID-19

By Ajibola Elias Maryam

OVERVIEW

Introduction

Dataset

Methodology

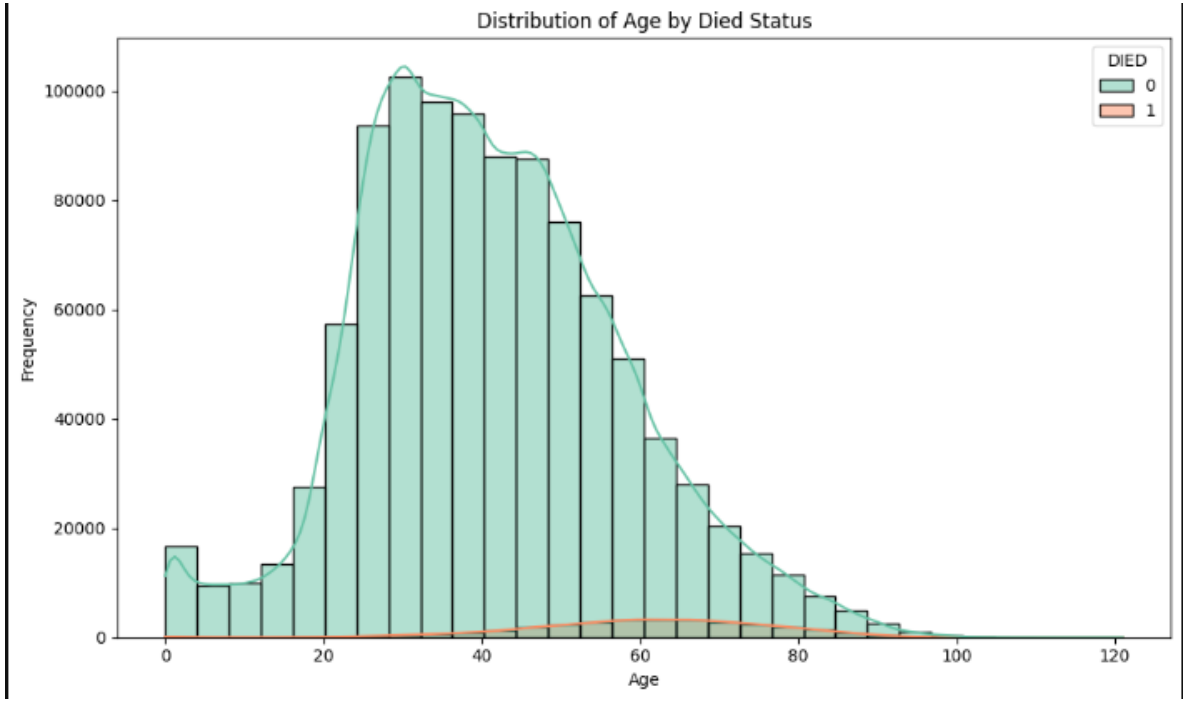
Outcome

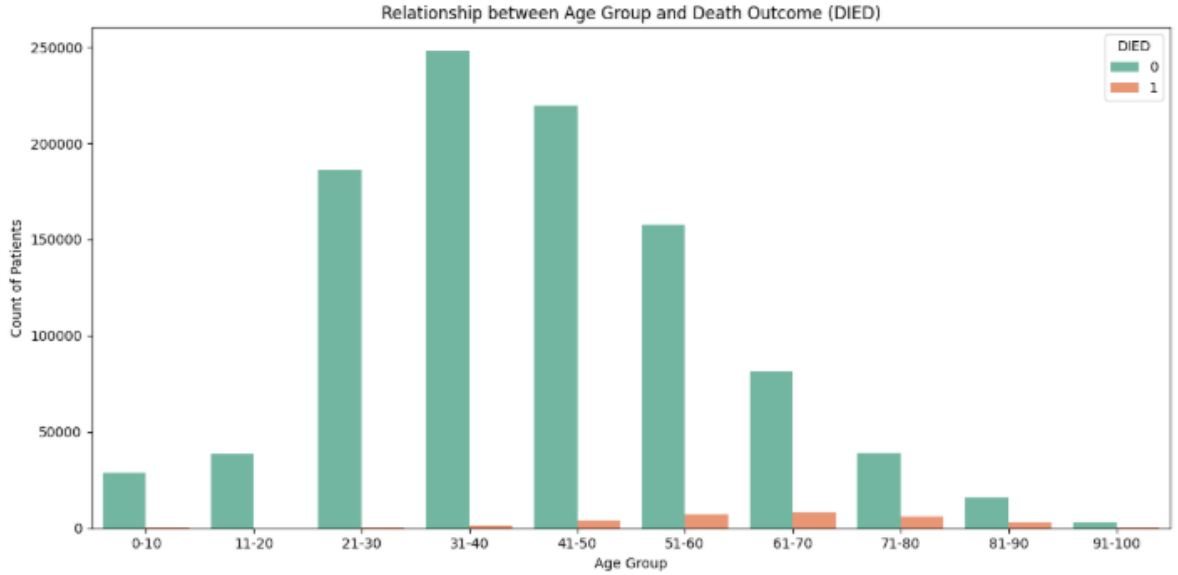
INTRODUCTION

Background Information

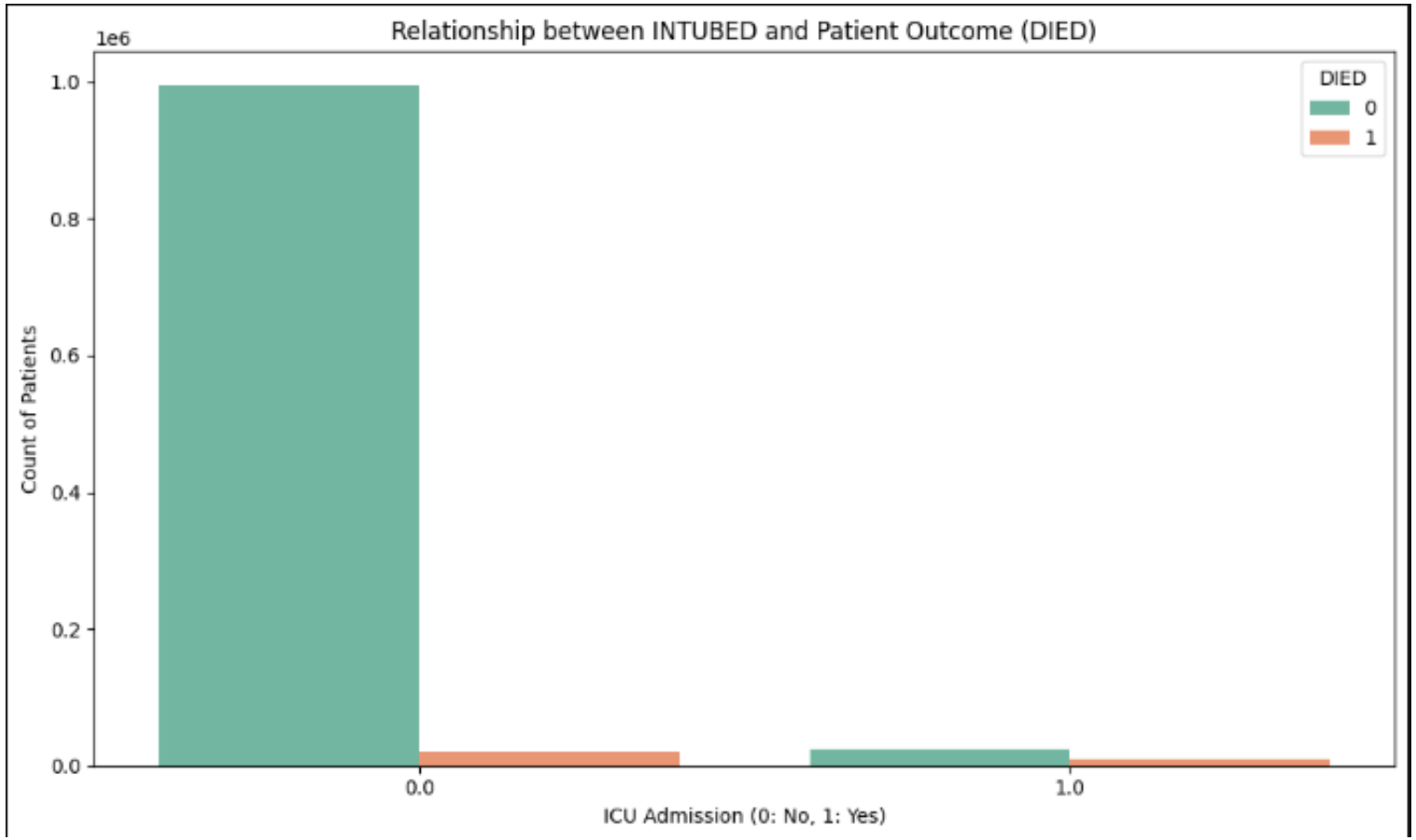
Objective

TARGET VARIABLE RELATIONSHIIP WITH OTHER FEATURE-AGE

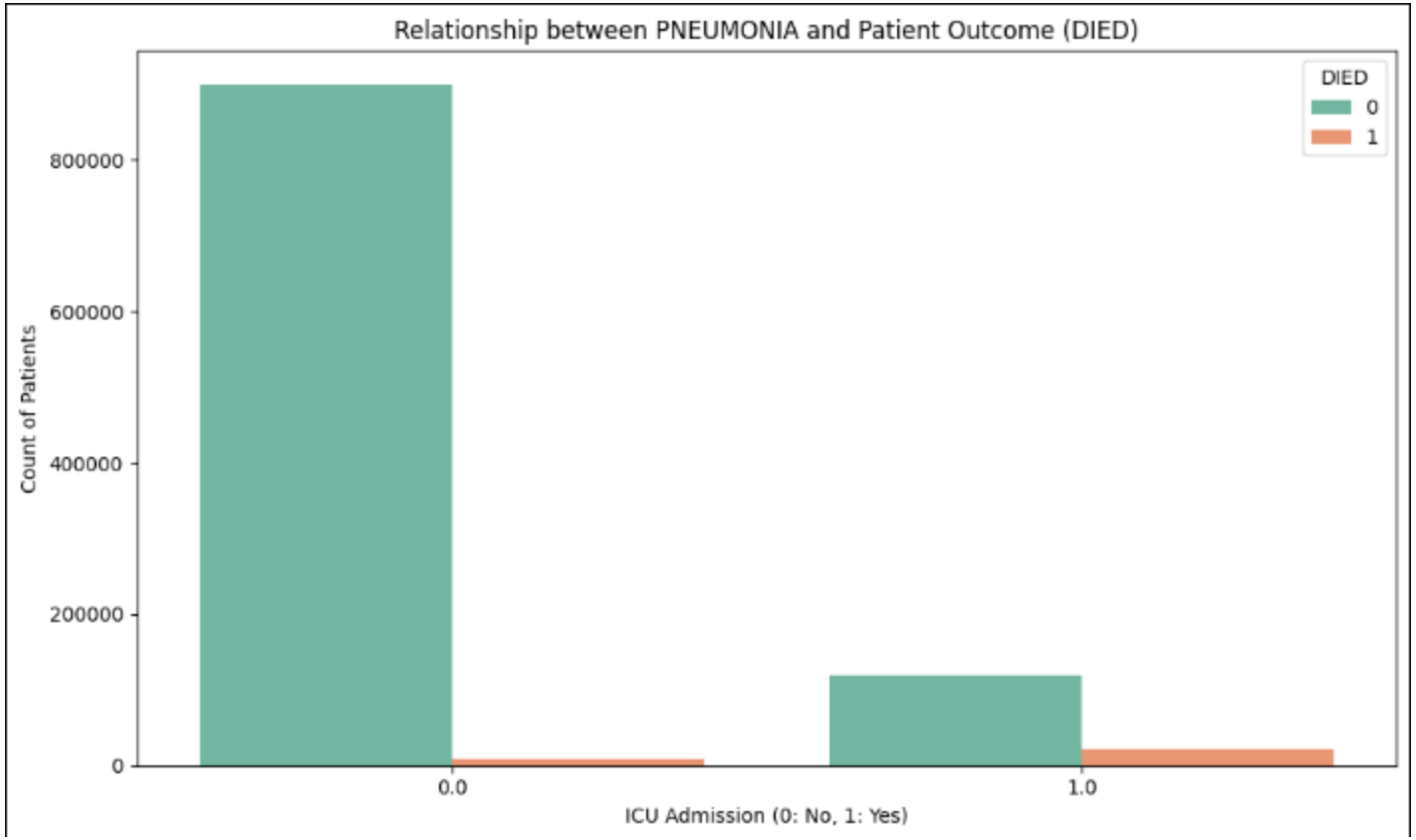




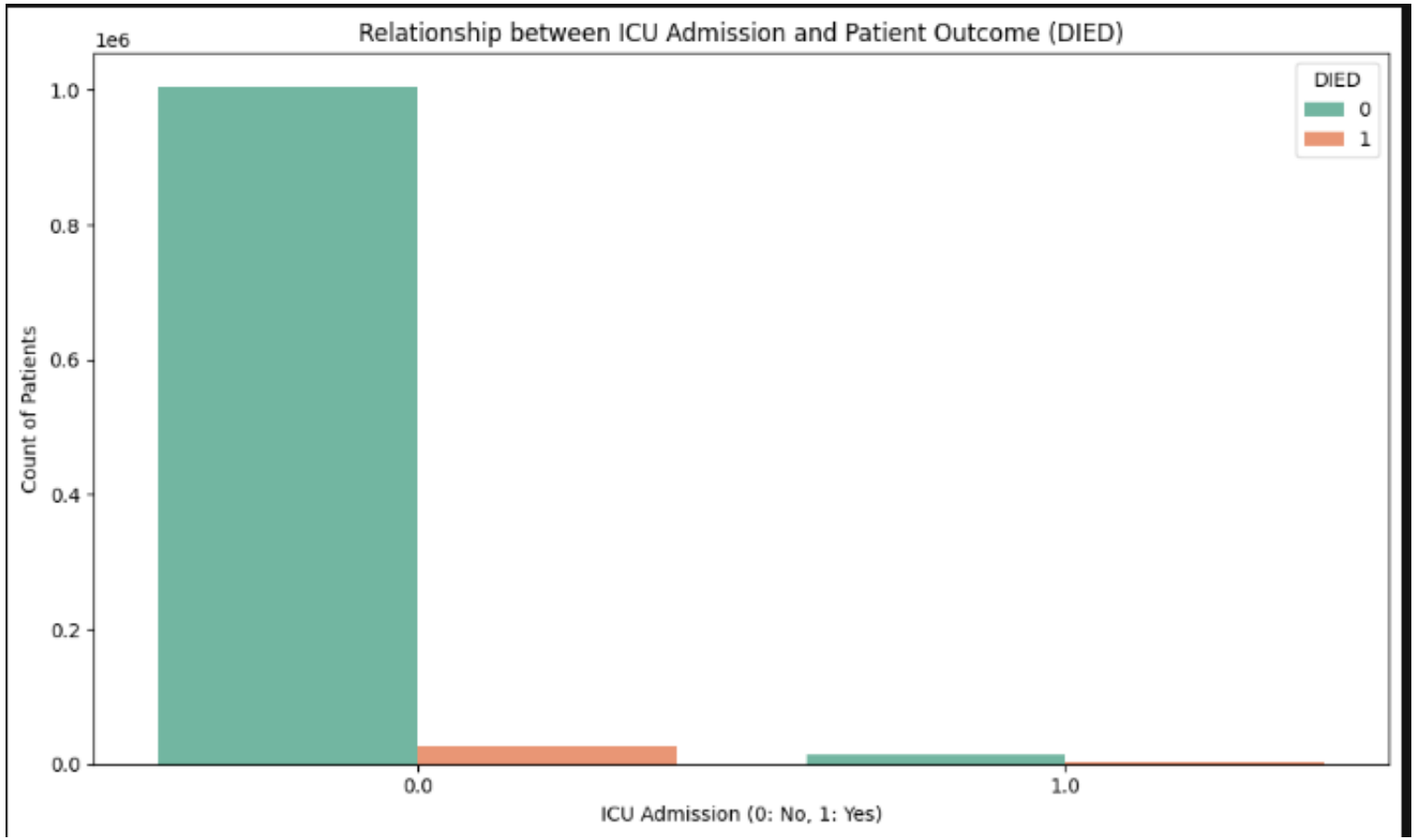
INTUBED V. DIED



PNEUMONIA V. DIED



ICU ADMISSION V. DIED

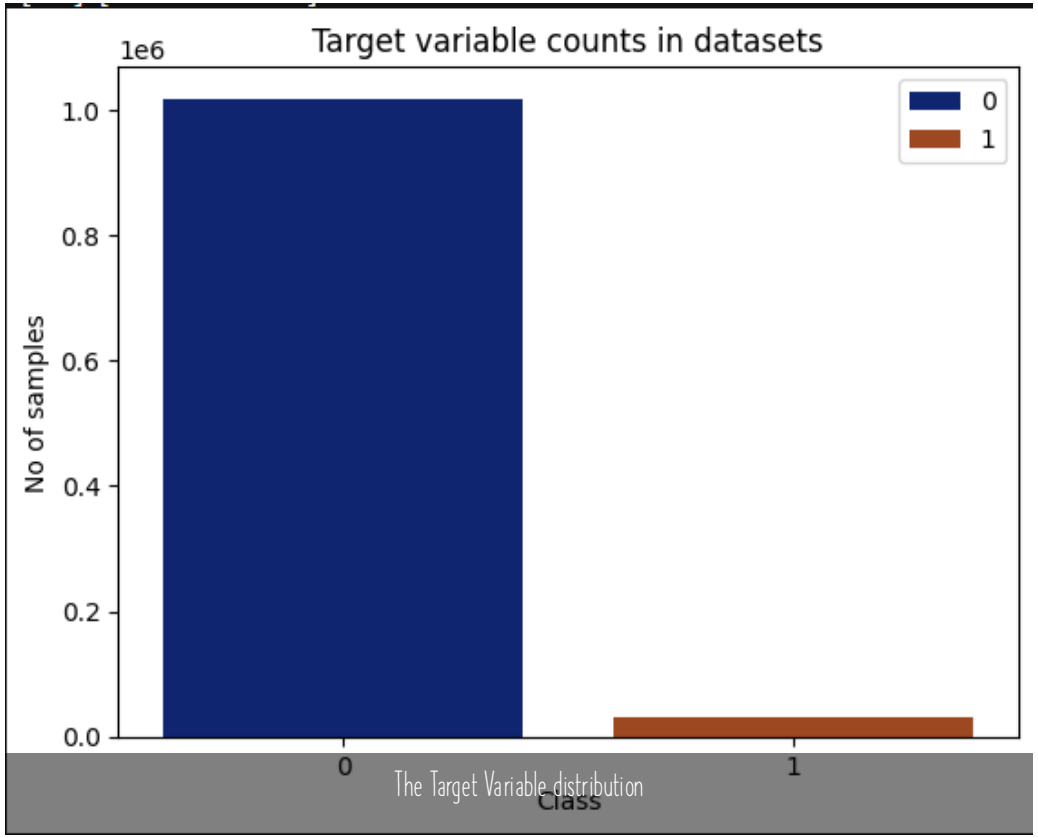


METHODOLOGY

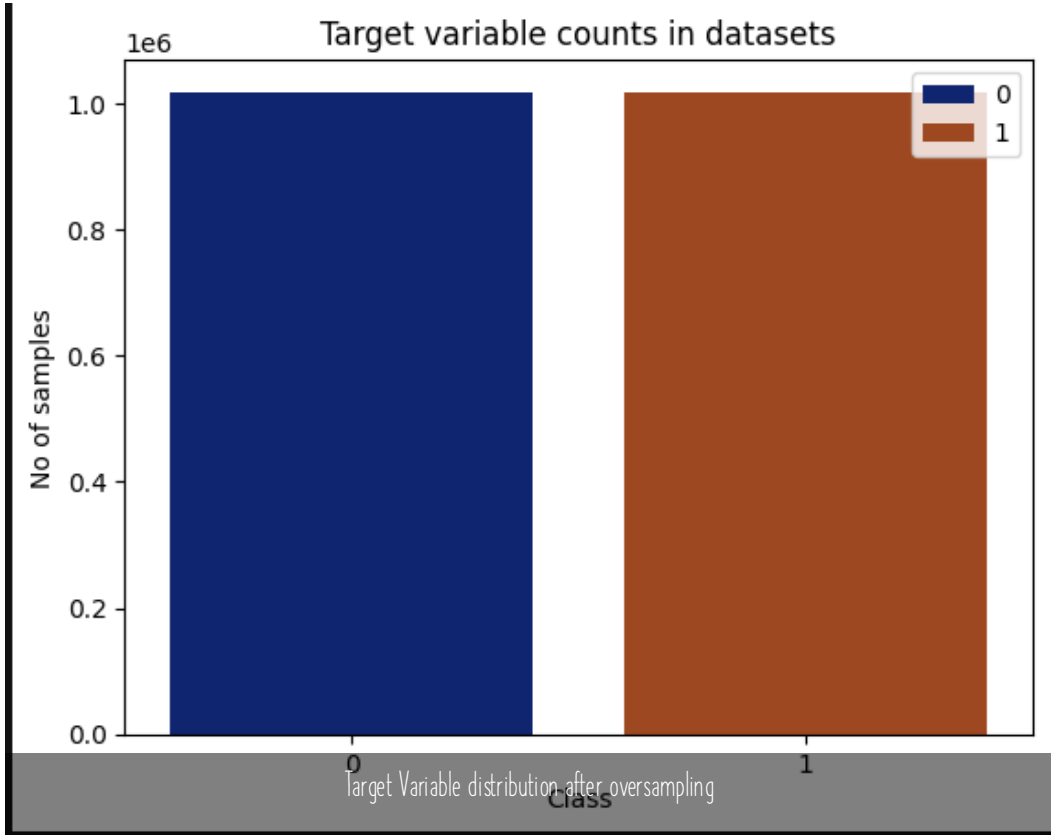
Since this is dataset doesn't have a defined target variable, I have decided to make one of the columns 'DATE_DIED" as my target variable.. For the model, I will be doing XGBoost and Random Forest since their good for a binary classification dataset. I'm going to explore using a Neural Networks like FNN, since though the dataset is quite complex.

Model Evaluation: Confusion matrix, F1-Score, Accuracy, Precision, & Recall. Also the AUC-ROC.

BALANCING THE DATASET



The Target Variable distribution



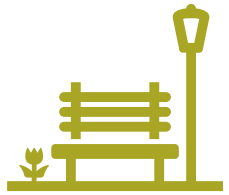
Target Variable distribution after oversampling

MODELS

XGBoost:

Pro's: Robustness -Resistant to overfitting; High Performance, and Feature Importance.

Cons: Computationally Expensive & Complex Hyperparameter Tuning,



Random Forest:

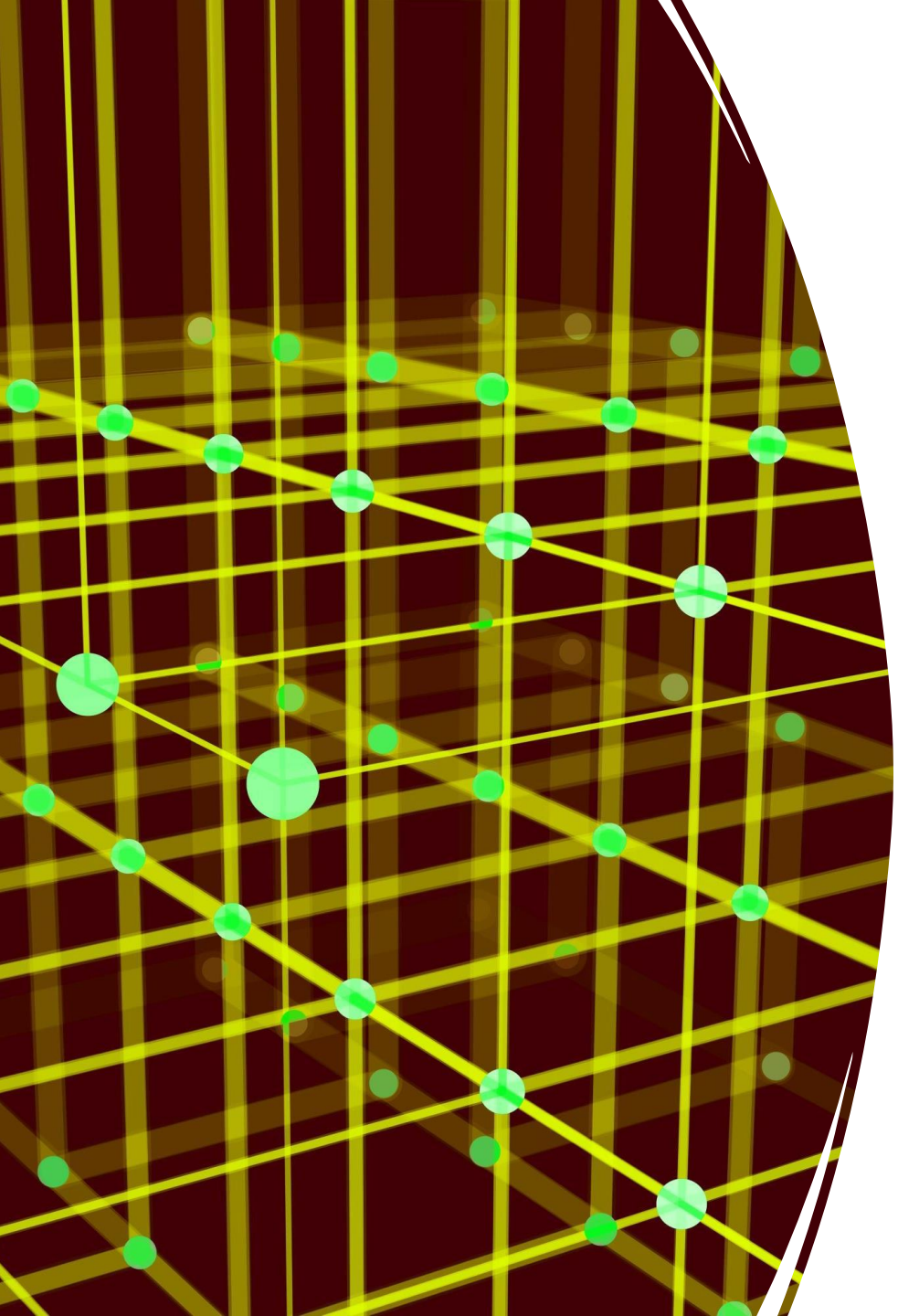
Pros: Handles Overfitting, Feature Importance, Robust to Noisy Data.

Cons: ComputationIlly Expensive, &Bias toward High Frequency Data.

FNN:

Pros: Flexibility & Versatility, Scalability, & Feature Engineering Reduction.

Cons: Computationally Expensive



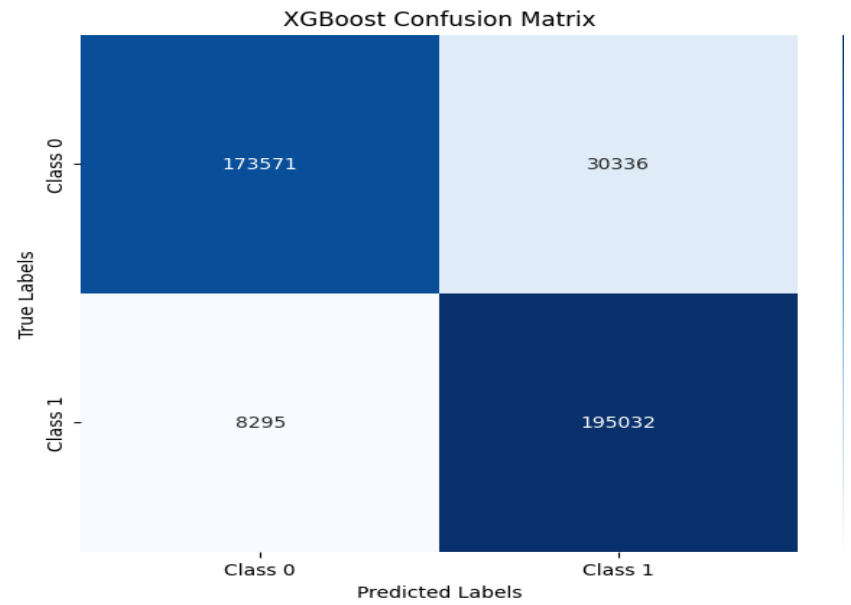
OUTCOME

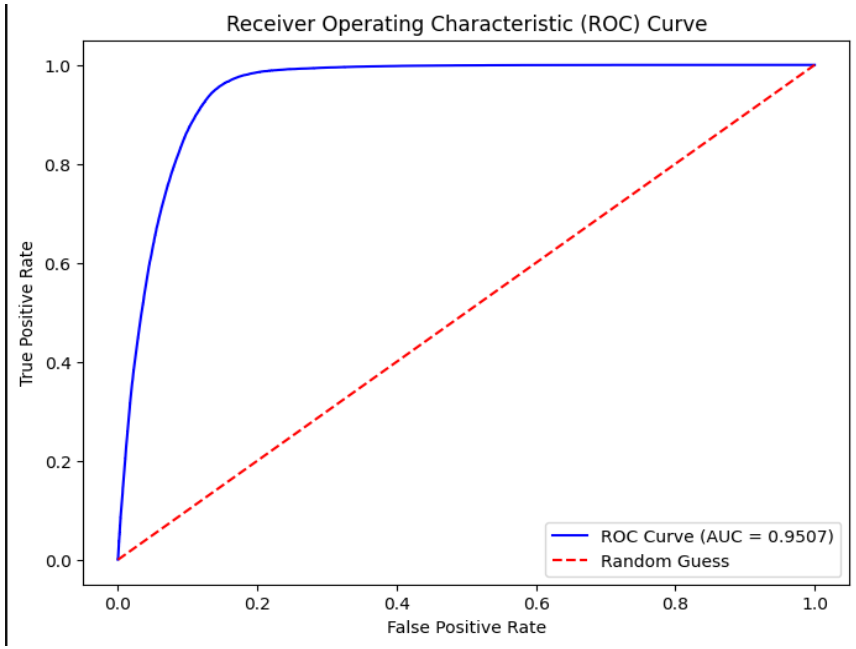
XGBOOST MODEL

- Evaluation Metrics:

```
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.85      0.90    203907
           1       0.87      0.96      0.91    203327

    accuracy                           0.91    407234
   macro avg       0.91      0.91      0.90    407234
weighted avg       0.91      0.91      0.90    407234


Accuracy: 0.9051
Precision: 0.8654
Recall: 0.9592
F1-Score: 0.9099
ROC AUC Score: 0.9507
```

Confusion Matrix & ROC-Curve



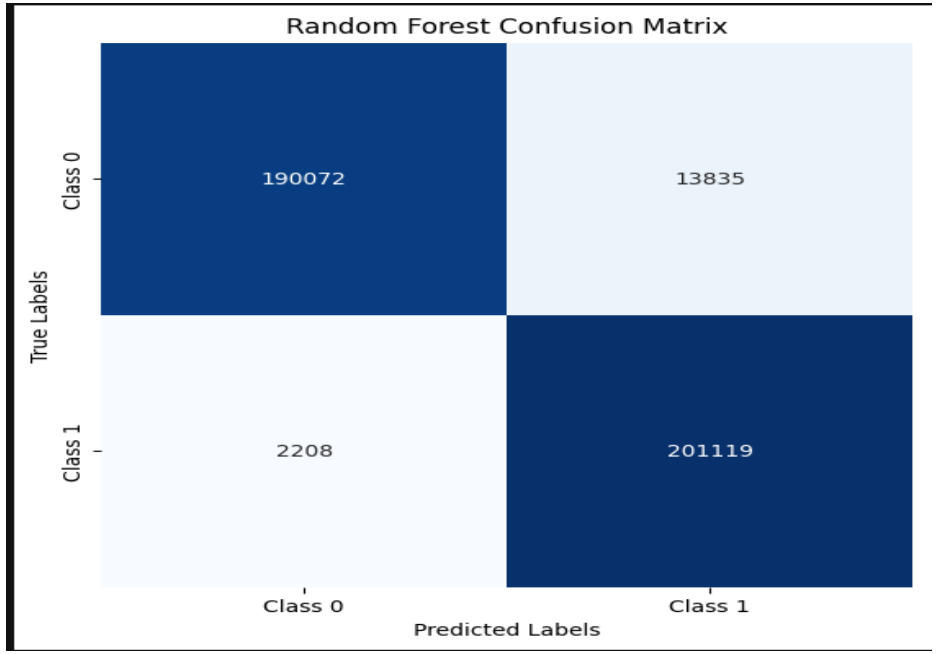


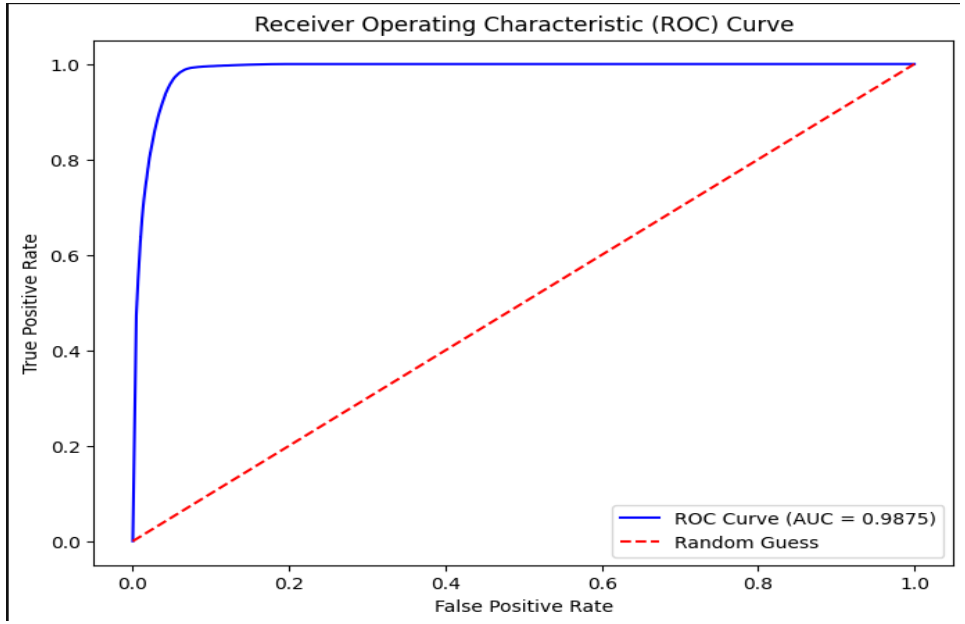
RANDOM FOREST

- Evaluation Metrics:

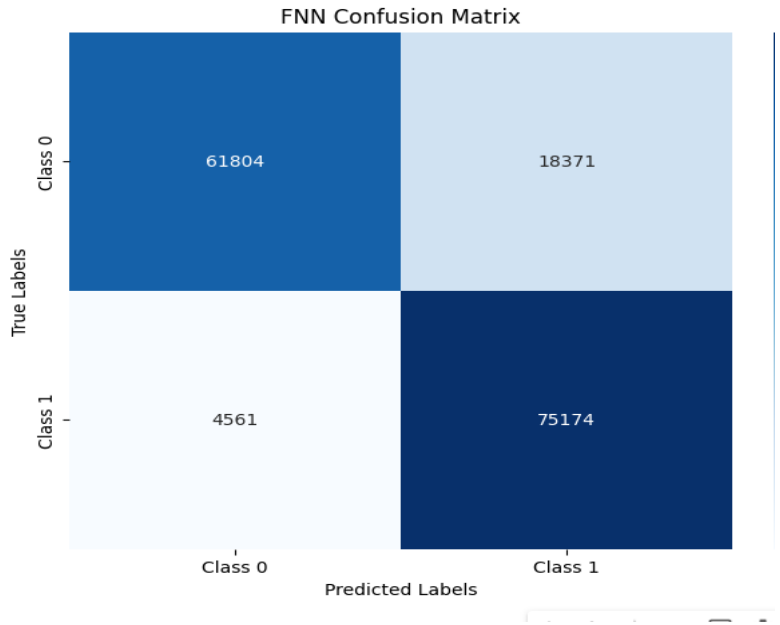
```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.93      0.96    203907
           1       0.94      0.99      0.96    203327

    accuracy                           0.96    407234
   macro avg       0.96      0.96      0.96    407234
weighted avg       0.96      0.96      0.96    407234


Accuracy: 0.9606
Precision: 0.9356
Recall: 0.9891
F1-Score: 0.9616
ROC AUC Score: 0.9875
```

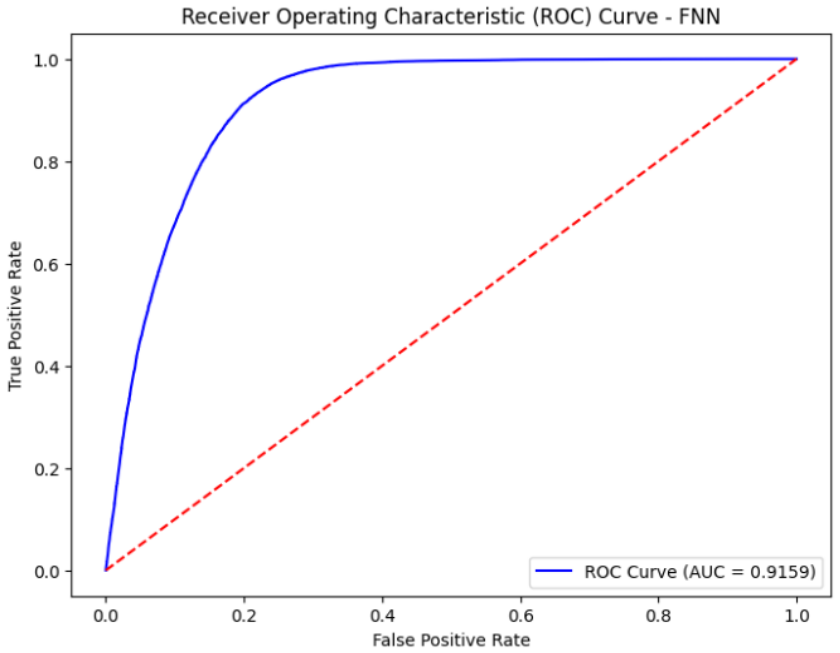
Confusion Matrix & ROC-Curve





Confusion Matrix & ROC_Curve





FNN(FEEDFORWARD NEURAL NETWORK)

- Evaluation Metrics:

```
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.77      0.84     80175
           1       0.80      0.94      0.87     79735

    accuracy                           0.86    159910
   macro avg       0.87      0.86      0.86    159910
weighted avg       0.87      0.86      0.86    159910

Accuracy: 0.8566
Precision: 0.8036
Recall: 0.9428
F1-Score: 0.8677
ROC AUC Score: 0.9159
```

CONCLUSION

- Random Forest had the best overall performance & evaluations metrics. While XGBoost & FNN performenced well, they did not match Random Forest metrics with high recall (0.98) and ROC-curve (0.98).
- Potential application in healthcare, since the objective is to create a model good at detecting high risk patient, so urgent care can be delivered swiftly, especially in the case of Covid-19, which is still affecting us today.

CHANLLENGES FACED

- **Missing Data**: This was my 1st working with a missing data being represent as something other than Nan, so I was a little stumped on what to do when first encounter, by replacing with something familiar(Nan), I was able to work everything out.
- **Date Imbalance**: In the data, there seem to be an overrepresentation of people who are not high risk. This makes it less obvious which feature has an influence on the target variables, to combat this issues this the data was balanced out by oversampling
- **Feature Selection**: Since the dataset was unsupervised, it was challenging to determine which column to use as the target variable while aligning with the objective. I initially attempted feature engineering to create a target variable, but this led to overfitting in most models. Taking a step back, I decided not to overcomplicate things. I chose 'DATE_DIED' as the target variable, as it most closely aligned with the objective of determining whether a patient was high risk, and it provided a clear indication of whether the patient survived or not.

PLANS FOR CAPSTONE 2

- My plans for Capstone 2 is look for another dataset to work with. I would like to do Computer Vision Project. Most of the Data I work with a Tabular. Preferably a dataset realting to business or medical.