# Predicting Patient Survival

Brinda Asuri, Pranav Garg, Soham Bidyadhar, Grayson Merritt, Tom Starkie

# Meet the Team

Grayson Merritt

Pranav Garg

Tom Starkie

Soham Bidyadhar

Brinda Asuri

# Overview of Content

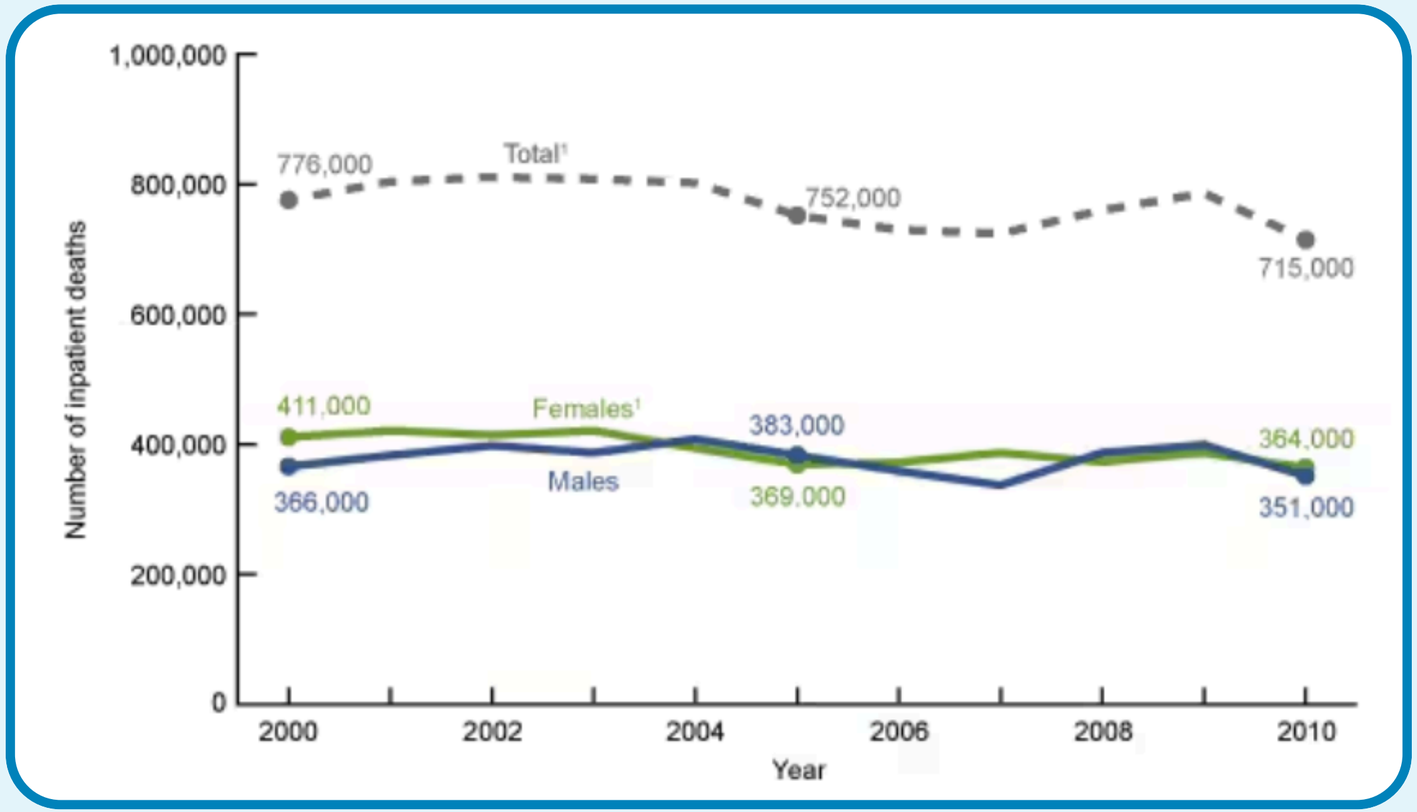


# 01

# Introduction

# Patient Survival Problem

- Accurate mortality predictions are essential for timely medical interventions and optimal resource allocation to enhance patient care
- Mortality risks vary across ICUs due to differences in staffing, resources, and local healthcare practices.
- **Existing research lacks globally applicable severity scoring systems that generalize well across diverse ICU datasets**
- Traditional severity scoring systems like APACHE are widely used but show calibration issues across different healthcare settings
- Despite an 8% decrease in patient hospital deaths from 2000 to 2010, over 700,000 patients died in U.S. hospitals in 2010




# Project Goal

## Develop a robust machine learning model to accurately predict hospital deaths using comprehensive international critical care data

Create a universally applicable severity of illness scoring system that performs reliably across diverse healthcare settings and countries

# 02

# Data Collection & EDA

# Data Collection

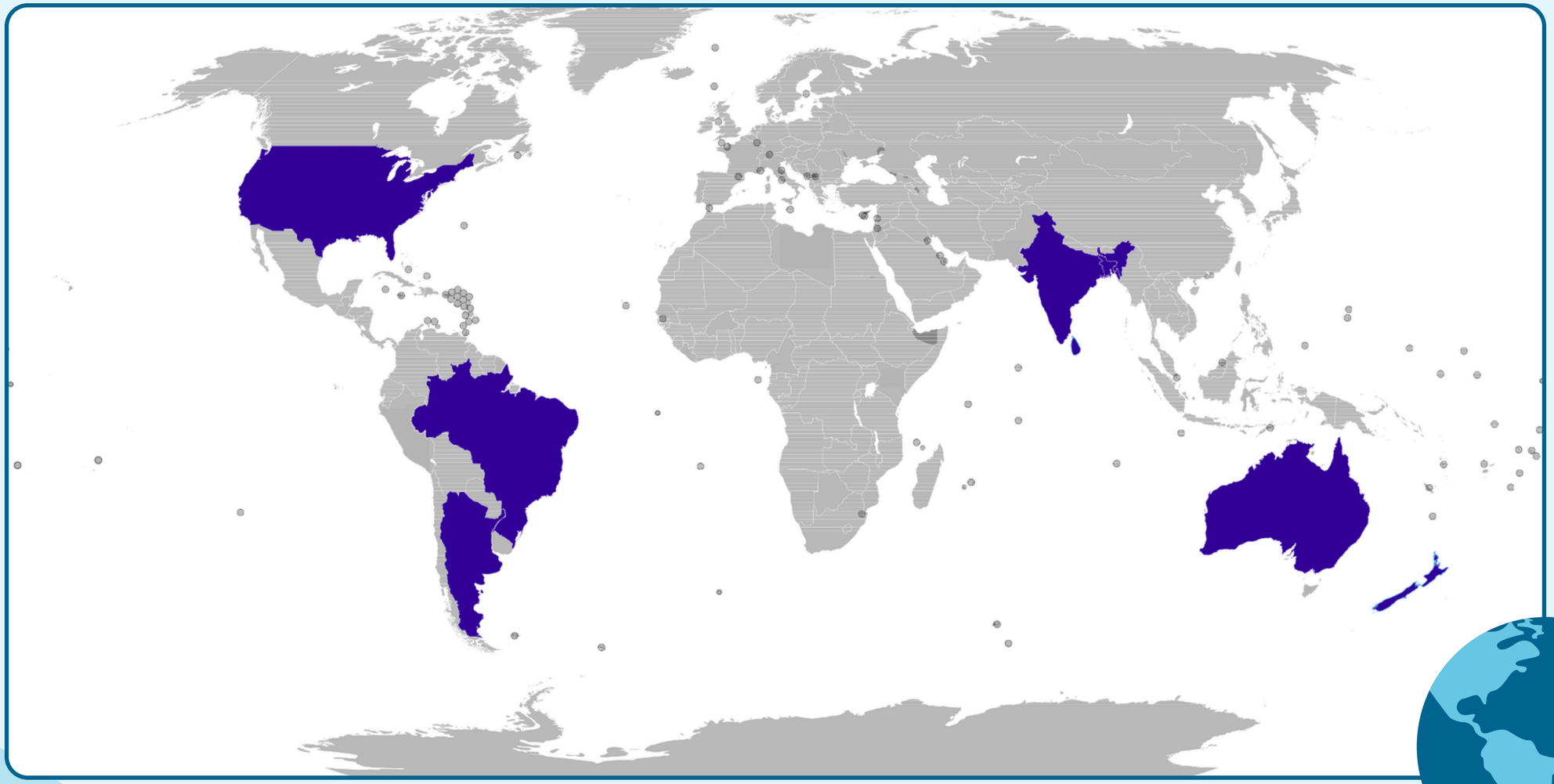
## Where does our data come from?

- **Data comes from the GOSSIS Consortium members** from Argentina, Australia, New Zealand, Bangladesh, India, Nepal, Sri Lanka, Brazil, and the United States
- Pooled critical care data from over **385 hospitals, encompassing 380,280 patients** across multiple countries
- A subset of patient data made its way to Kaggle - **91,713 rows**

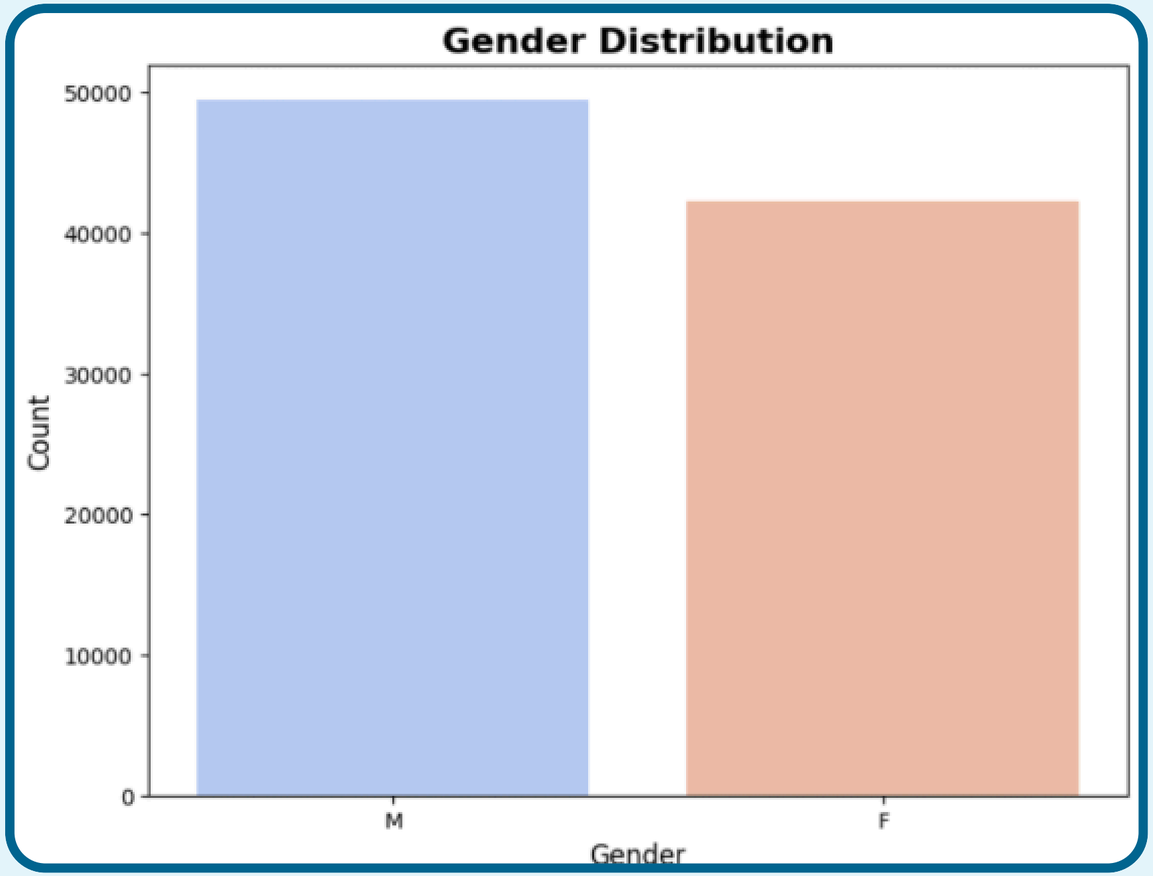
## What type of data was collected?

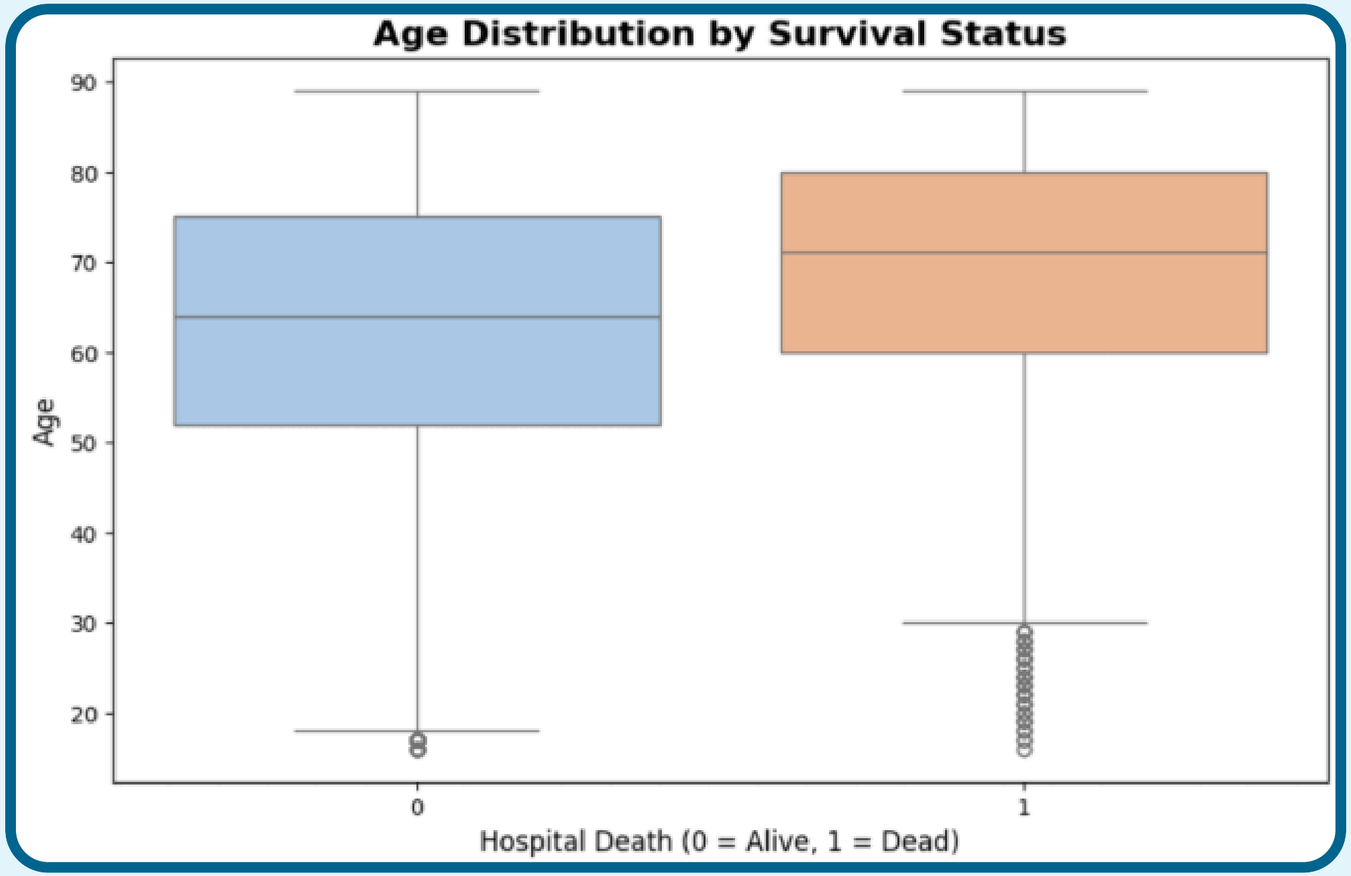
- **Collected comprehensive patient information,** including vital signs, laboratory results, admission diagnoses, Glasgow Coma Scale scores, chronic comorbidities, and demographic variables
- Focused on **data gathered within the first 24 hours of ICU admission** to capture early indicators of illness severity

# Countries in Dataset

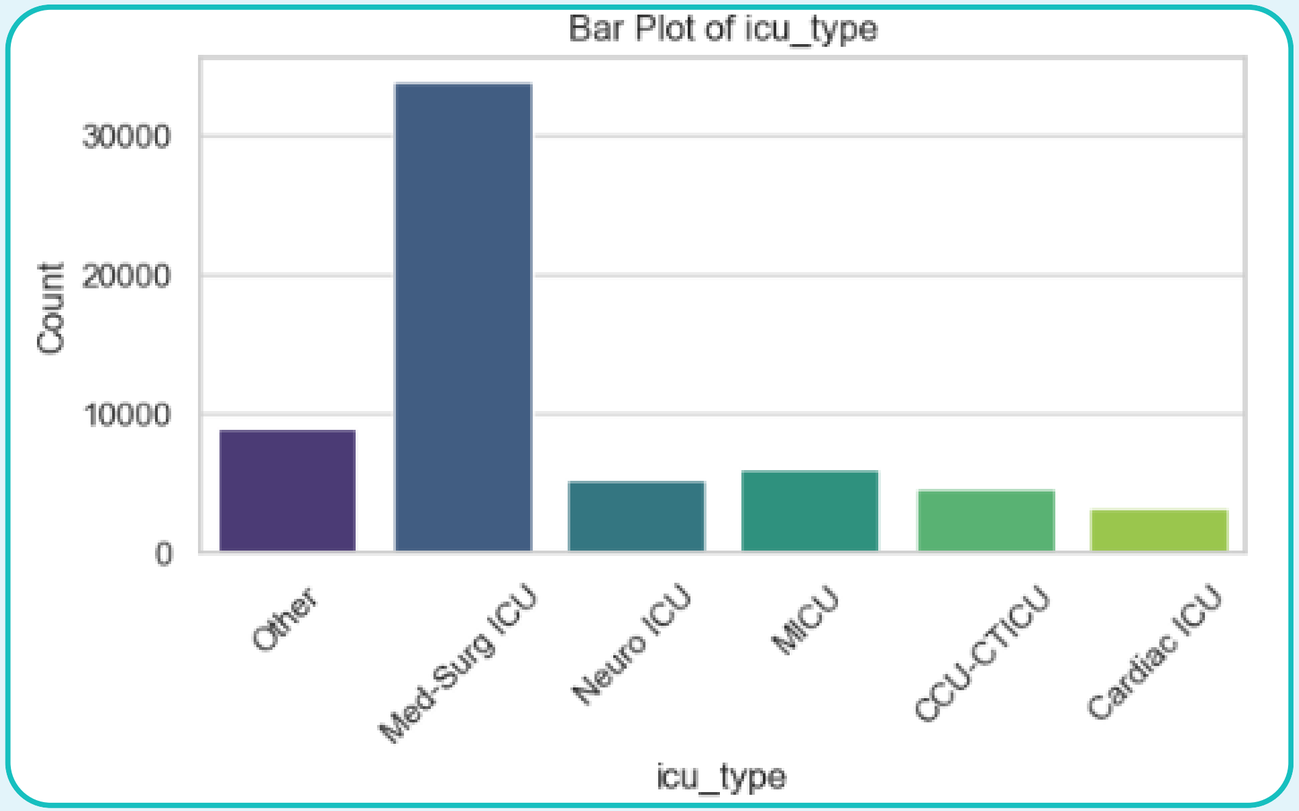


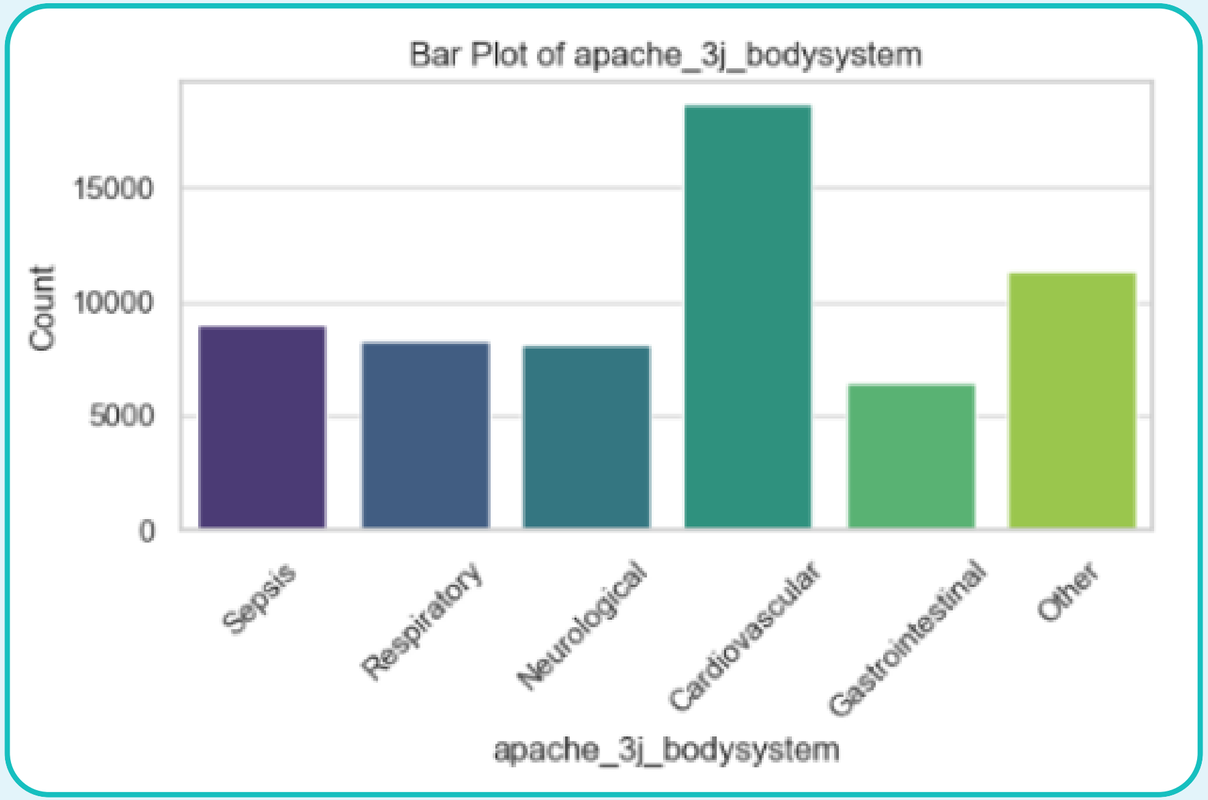
# EDA - Age and Gender Distribution



# EDA - Categorical Features



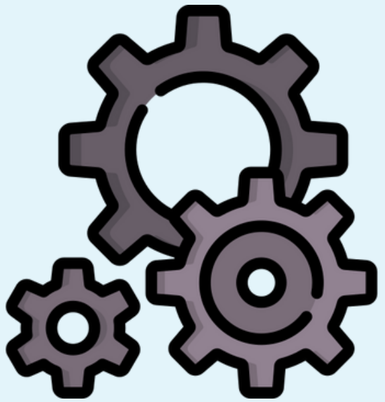
# 03

# Data Pre-Processing

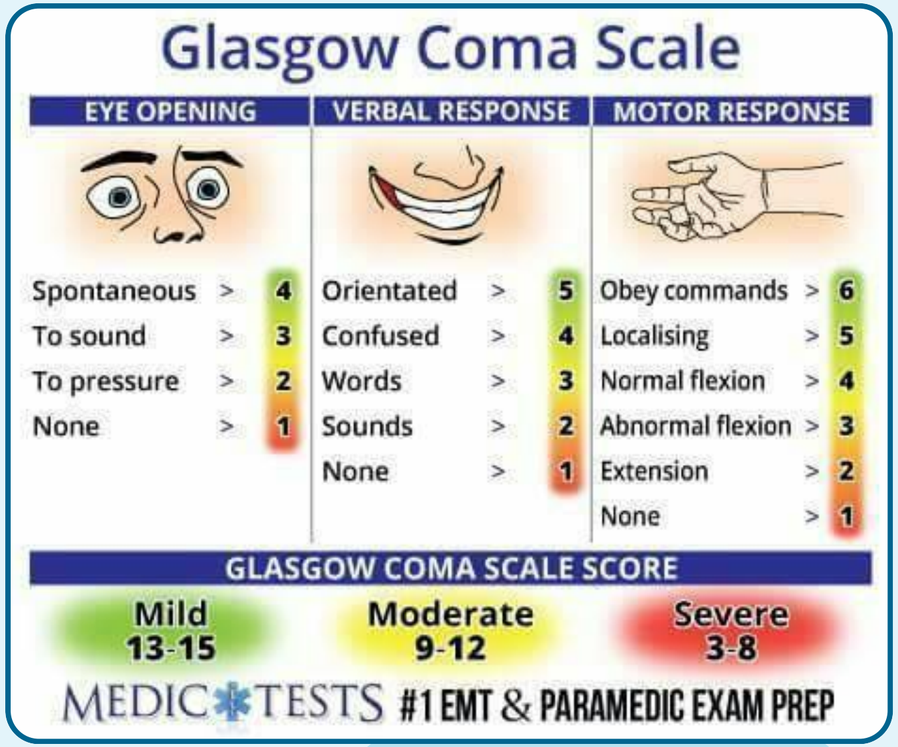
# Data Cleaning

- **Dropped** ID and blank columns
- **Save APACHE predictions** for comparison later
- **Missing values were random** - dropped them to keep things simpler
- **Experimented with imputation** using KNN - more time, same results

# Data Feature Engineering

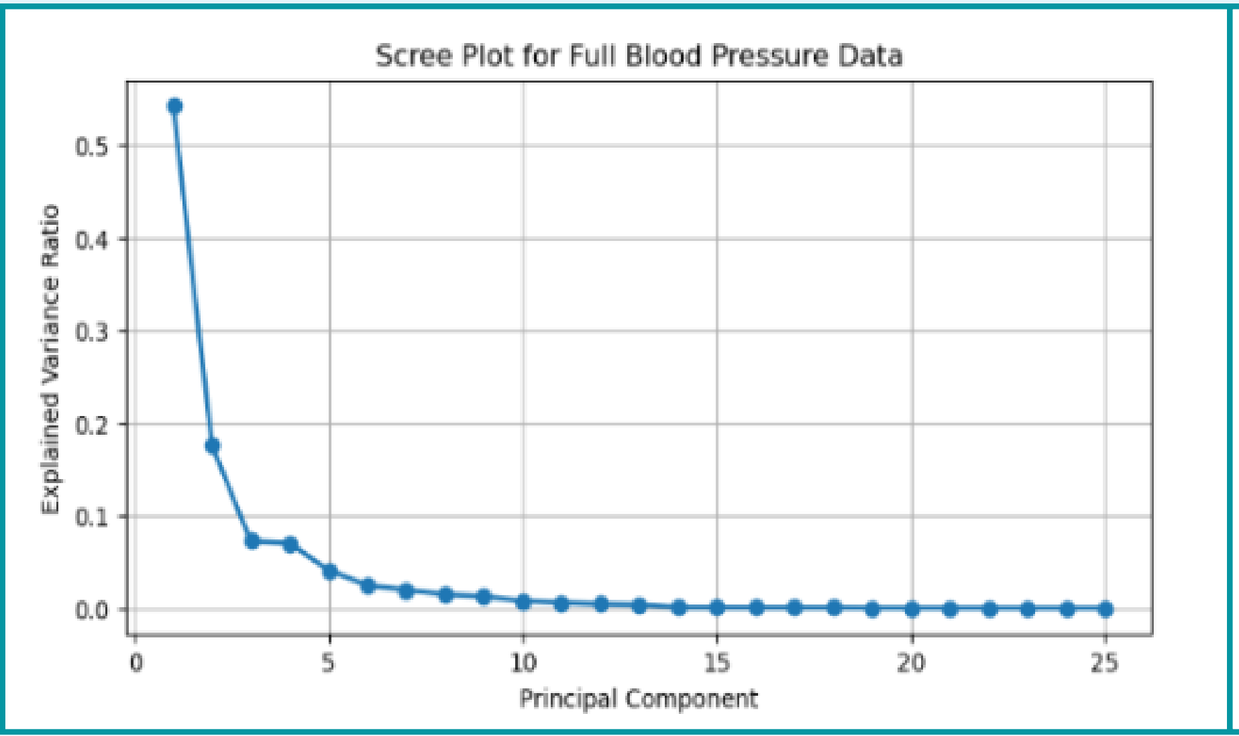


- Initial feature interactions did not add much to metrics
    - + want to avoid curse of dimensionality
- Collapsed Comorbidity and Glasgow Coma Scale features into a total variable for each
- Unimportant categories that are less than 10% are collapsed into “other”.
    - Ex. - “SICU” into “Other”
- Scale using StandardScaler to ensure all features contribute equally for models sensitive to scale



# PCA to the Rescue!

- Lots of physiological variables, and we wanted to to reduce dimensionality
- Used PCA on each subgroup (blood pressure, heart rate, etc.)
- Elbow Scree Plot is 🔑




**83**

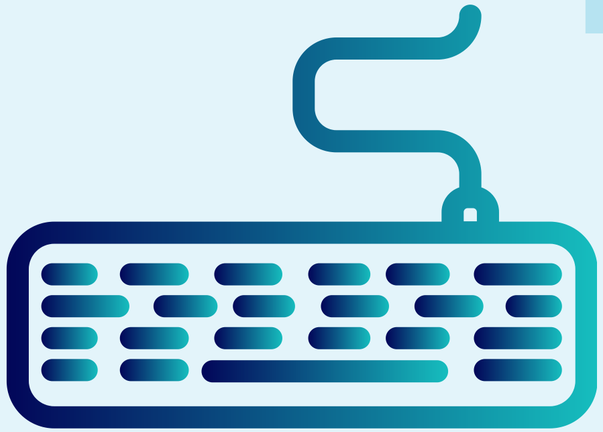
**Features**

**44**

**Features**

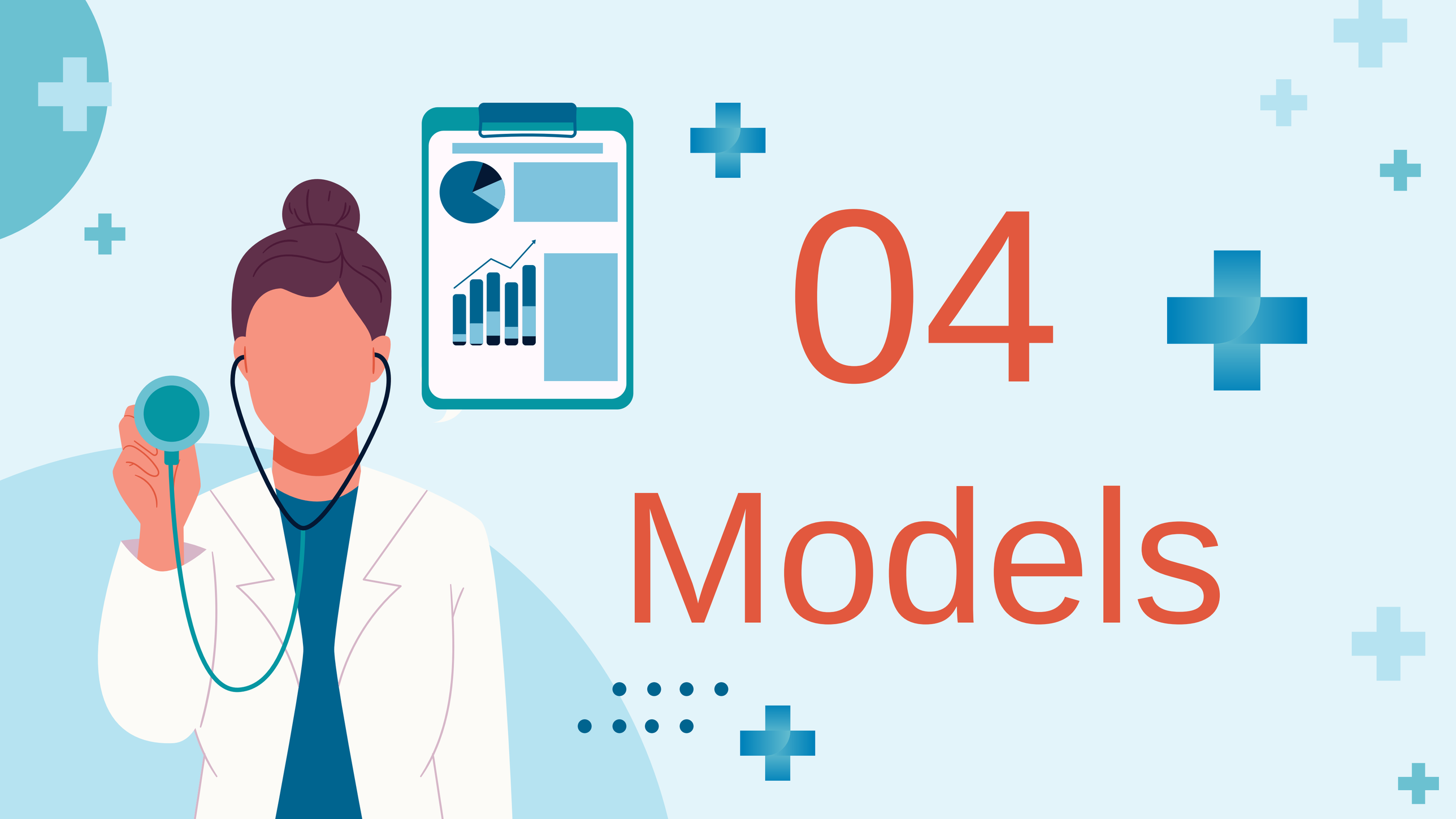
Dimensionality Reduced...

# Label Encoding Example



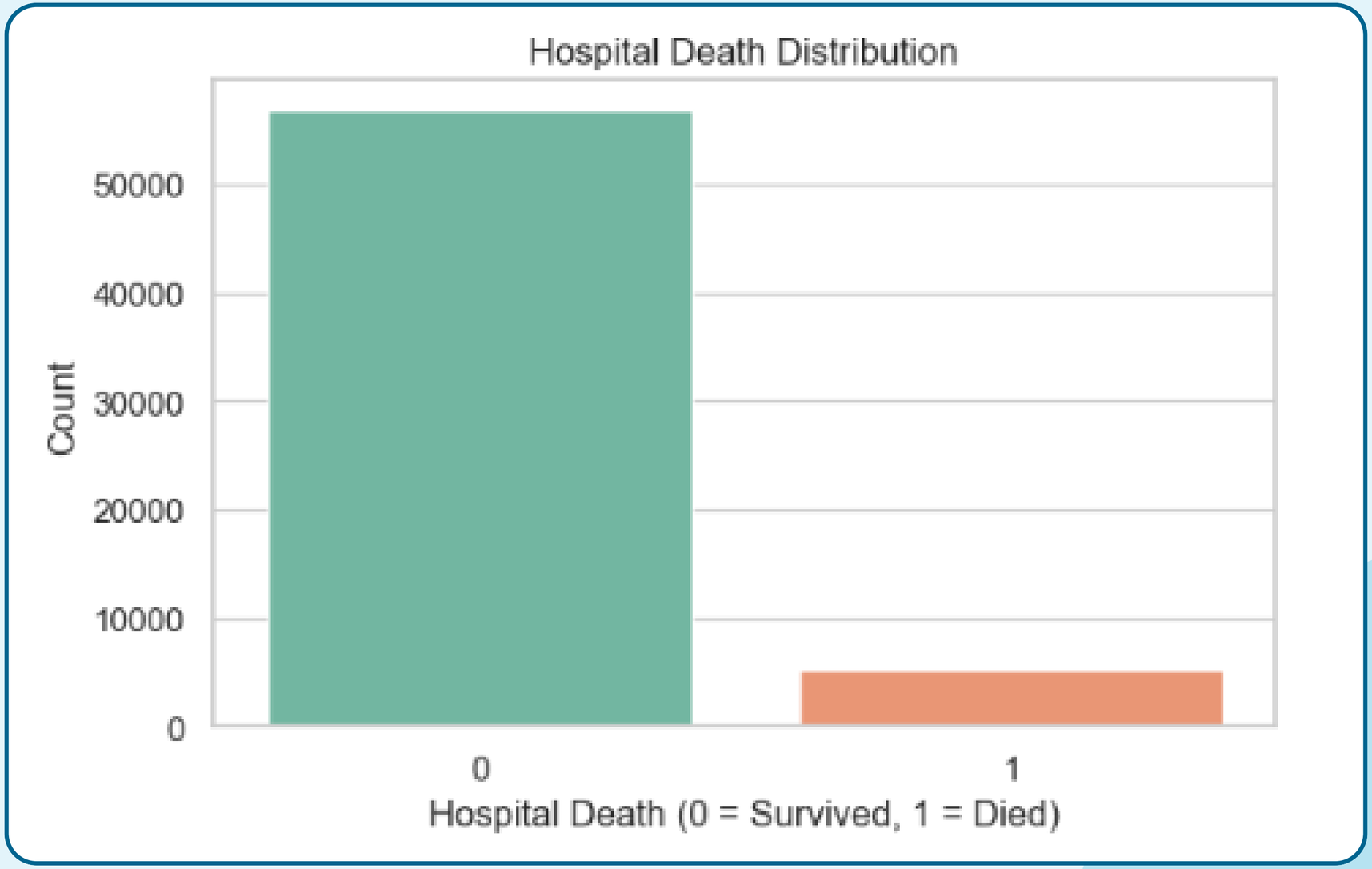
## ICU_stay_type

| | | |
|---|---|---|
| admit | → | 0 |
| re-admit | → | 1 |
| transfer | → | 2 |



# 04 Models

# We have a data problem...



# Data Balancing

## STRATEGY 1: Do Nothing

**Pros:**

- **Simple**
- Preserves original data distribution

**Cons:**

- Model **bias** towards majority class
- Poor minority class metrics

## STRATEGY 2: Use SMOTENC

**Pros:**

- **Balances** class distribution
- Improves minority class performance

**Cons:**

- Increased computational load
- Potential introduction of **noise**

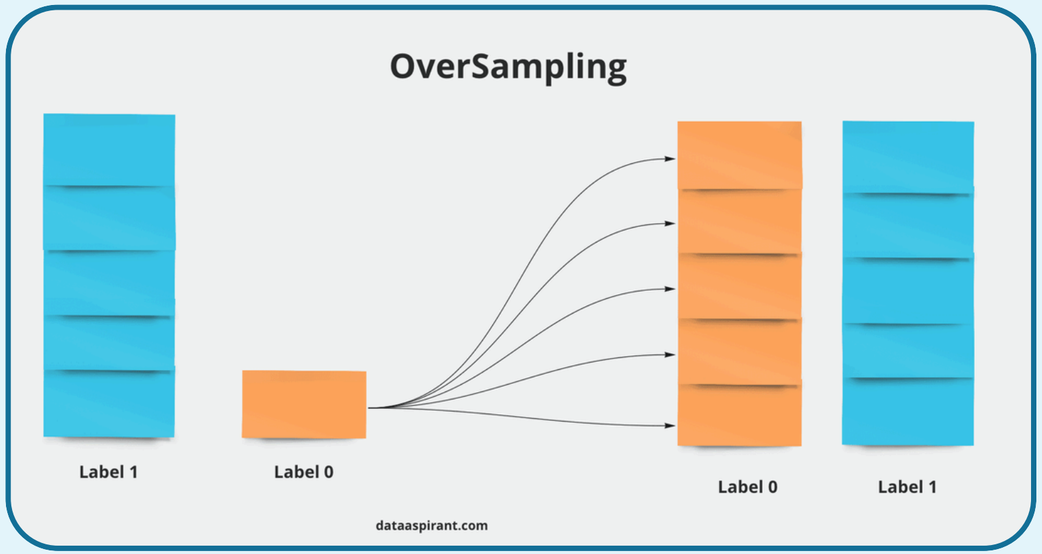
## STRATEGY 3: Adjust Class Weights

**Pros:**

- **Address imbalance** without changing data size
- Encourages model to focus on minority class

**Cons:**

- Less intuitive interpretability
- Potential **impact on majority class** performance



# A Smorgasbord of Models

9 model types * 3 strategies
= **27 models** to train

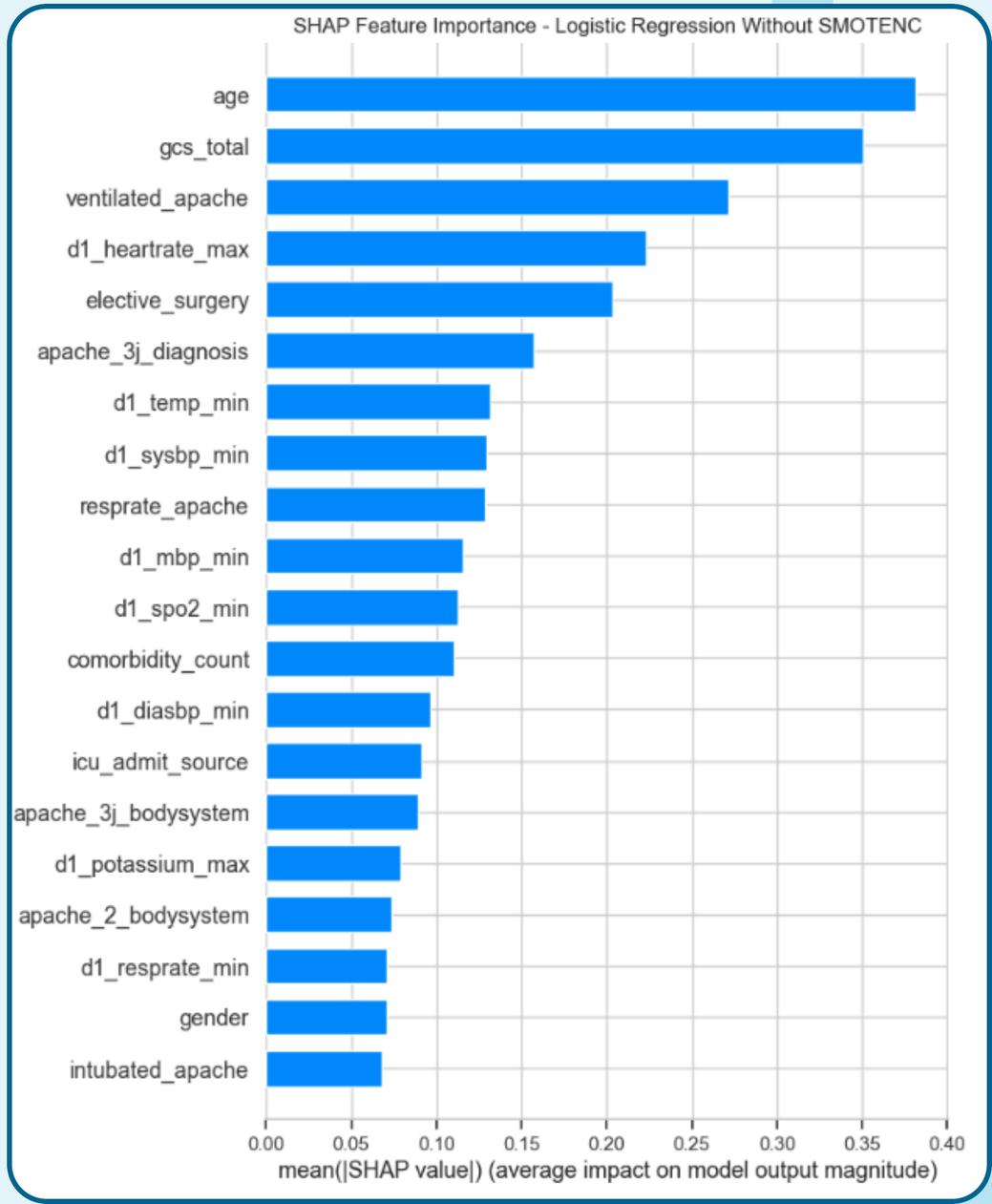
Bagging? ✓
Boosting? ✓
Naïve Bayes? ✓
Neural Nets? ✓
K Cross Fold Validation? ✓
Adjusting Thresholds? ✓
Hypertuning? ✗
Feature Importance? ✓

Crazy? ✓

# Logistic Regression: Class Weights

## Lasso Regression with a C of .1

**Accuracy:** .781
**Recall:** .766
**Precision:** .244
**F1-Score:** .370
**ROC AUC:** .854
**PR AUC:** .382



# Decision Tree - No SMOTENC

**Accuracy:** .874

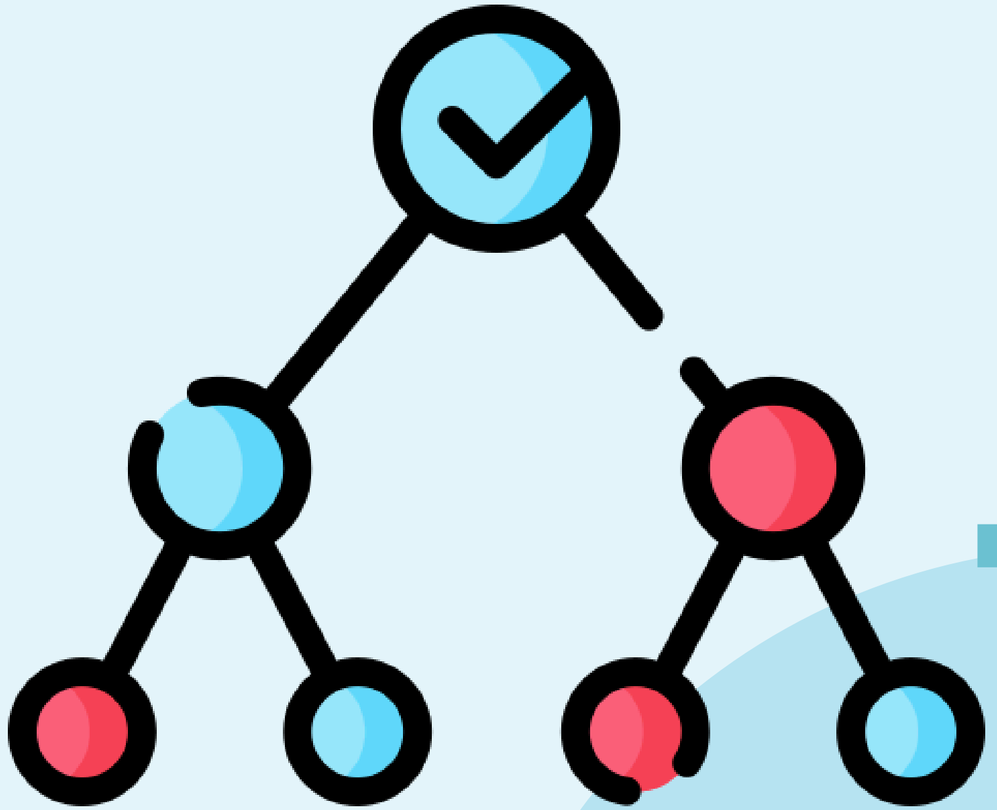
**Recall:** .311

**Precision:** .279

**F1-Score:** .386

**ROC AUC:** .618

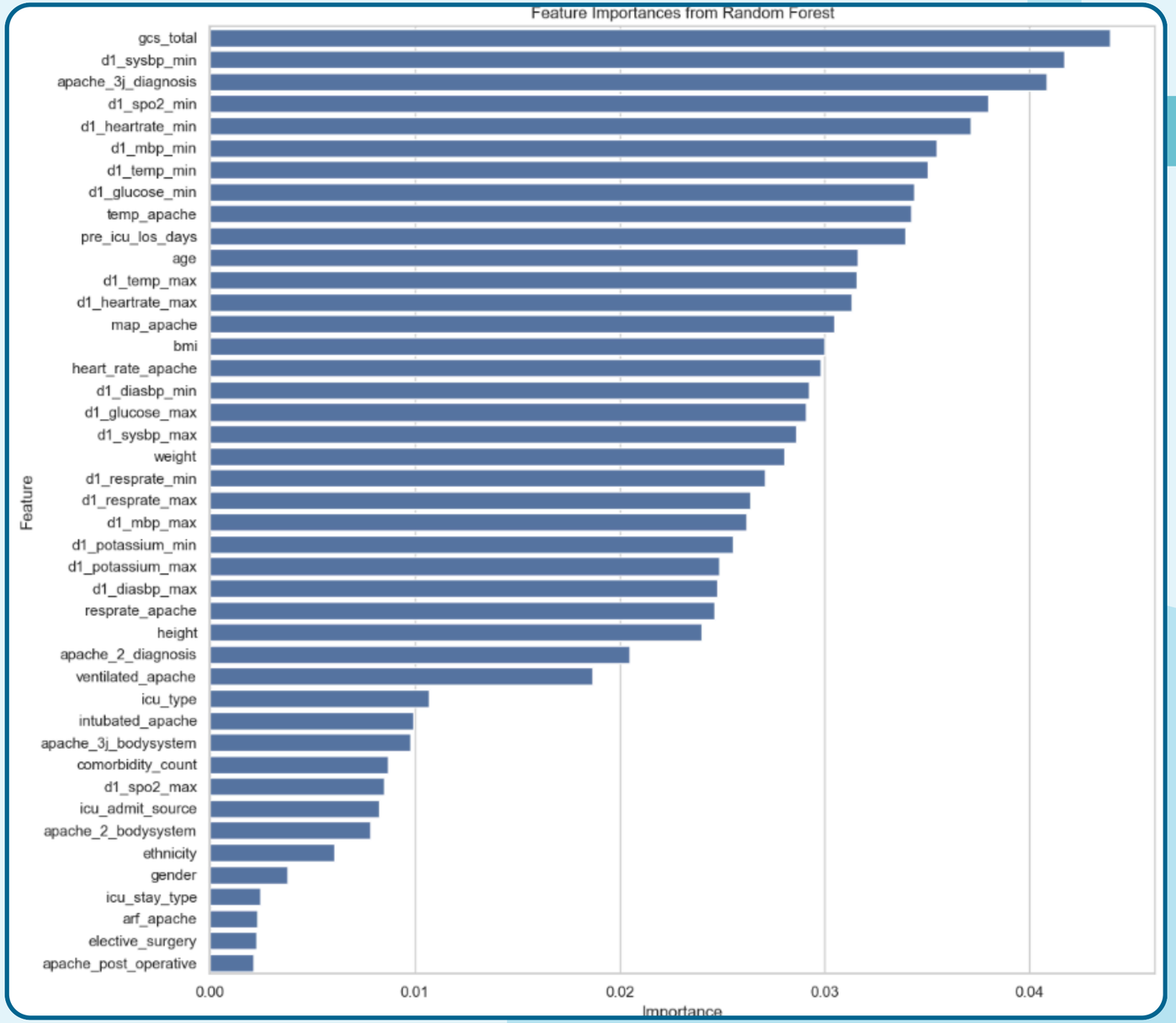
**PR AUC:** .324



# Random Forests: SMOTENC

**Accuracy:** .896
**Recall:** .386
**Precision:** .386
**F1-Score:** .386
**ROC AUC:** .849
**PR AUC:** .346

**Recall at 90%**
**Threshold:** .16
**Precision:** .176
**F1-Score:** .295



# Brief Intro to Threshold Adjusting

## OUR GOAL: Lower False Negatives

Threshold ⬇

Predicted Positive ⬆

Recall ⬆ (True Positive Rate)

Precision ⬇ (Increase False Positives)

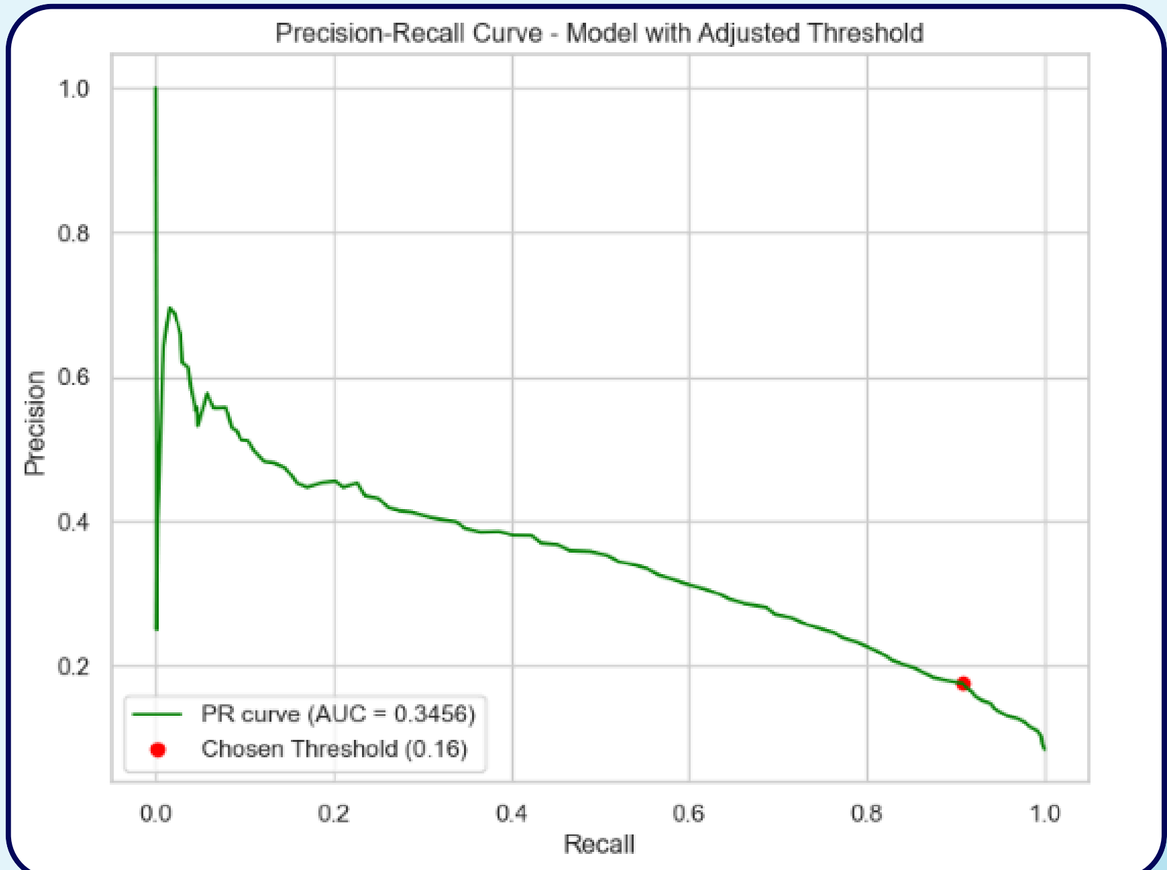
Accuracy ⬇

- Results in more false alarms, but can also provide more critical care for those who need it, ensuring that we do not miss any actual positive cases, even if it means allocating additional resources to review and manage these cases.

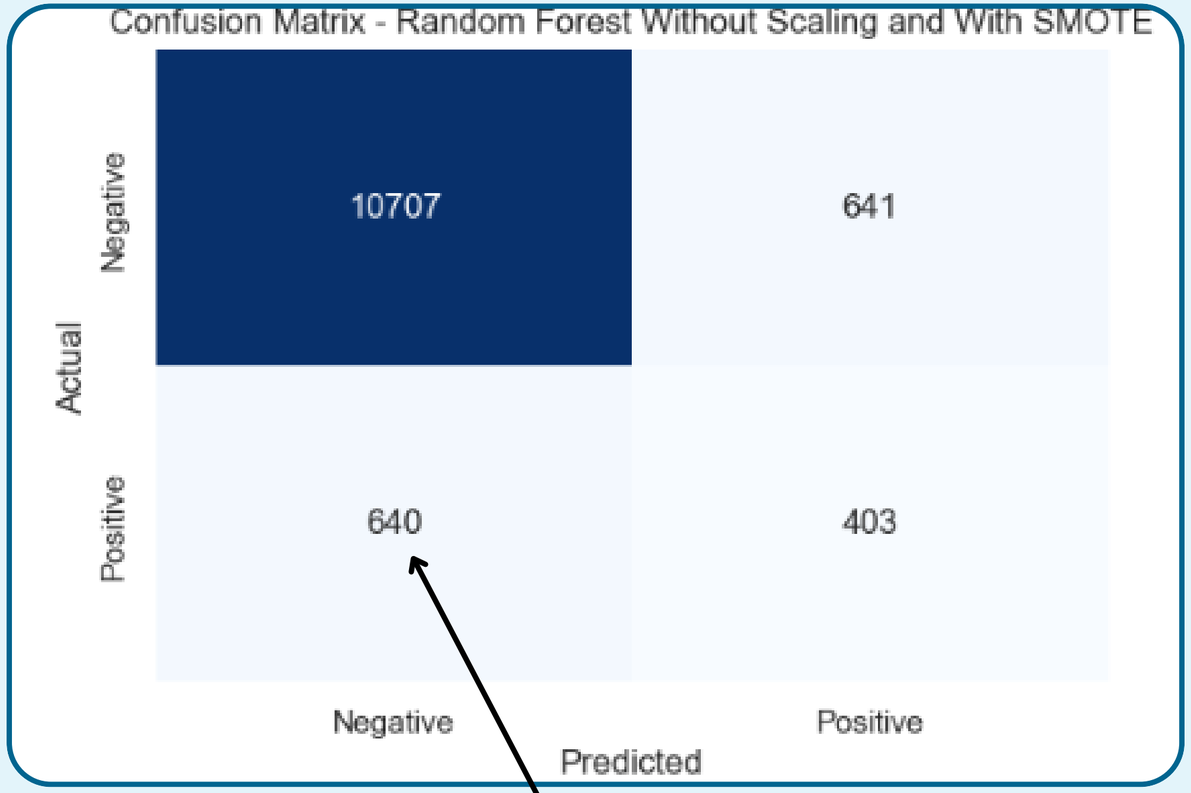
**Definition Reminders**

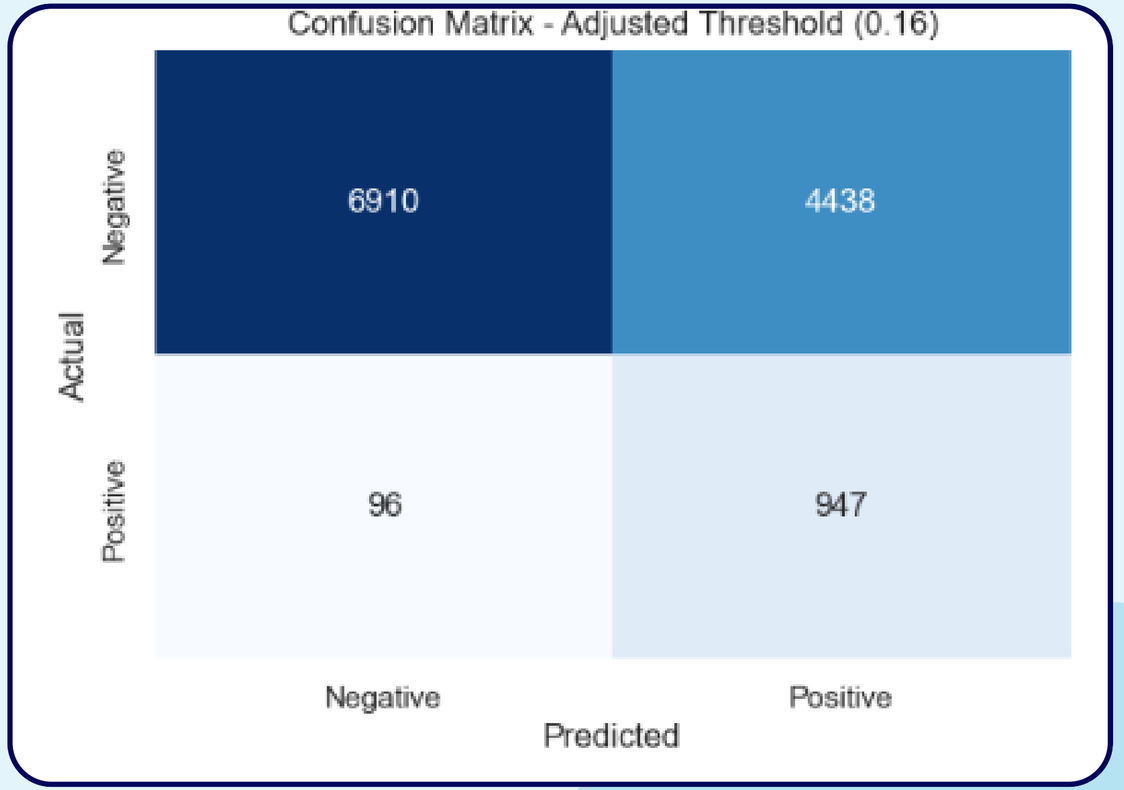
**Decision Threshold:** The probability cutoff point at which the model assigns a sample to a particular class. By default, this is often set at 0.5

**Recall for Class 1:** The ability of the model to correctly identify all actual positive instances (i.e., patients who will experience hospital death)



# Changing Thresholds Results



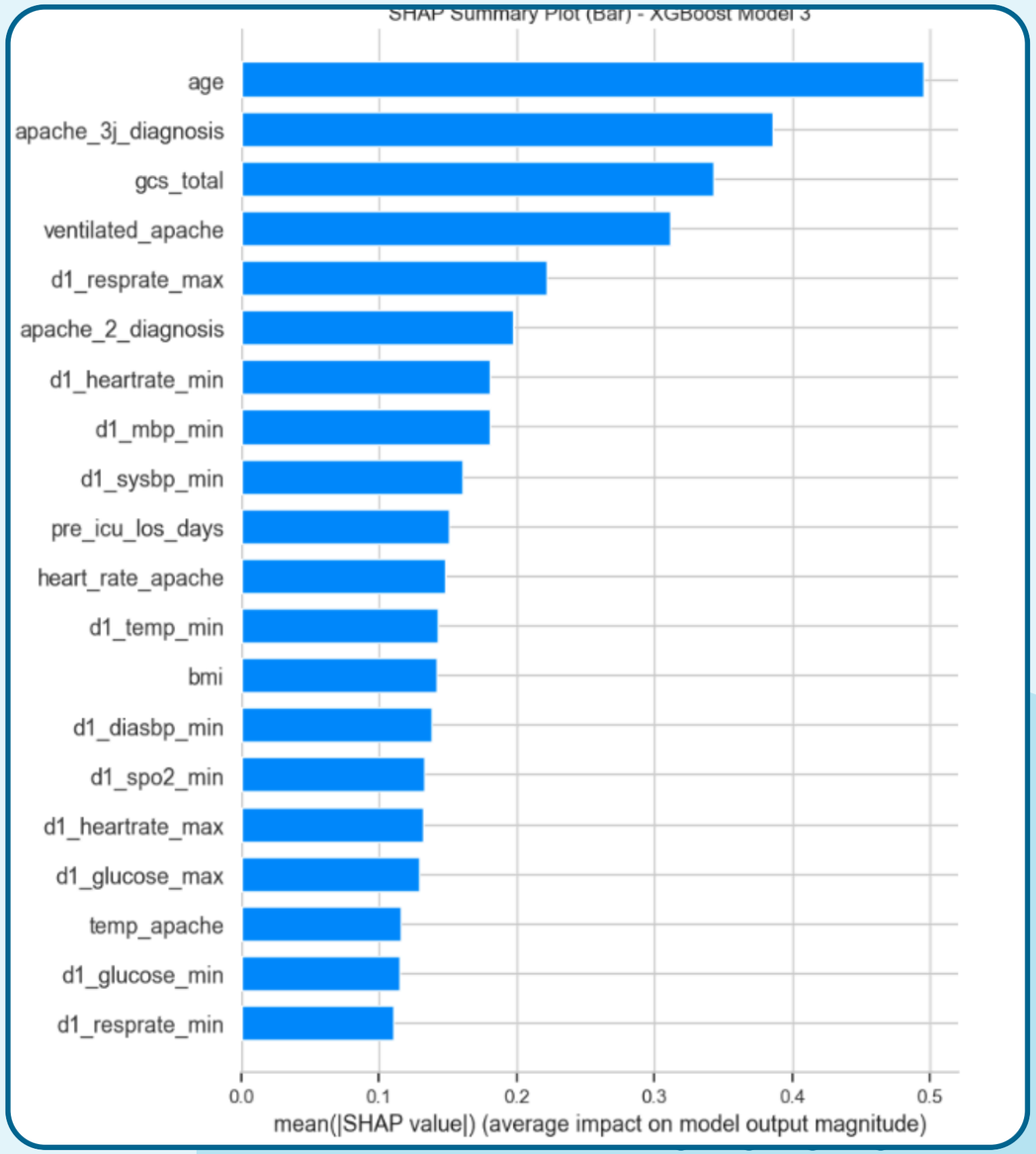

**Want this low**

**Shift to predicting more positives**

# XGBOOST: Class Weights

**Accuracy: .869**
**Recall: .585**
**Precision: .339**
**F1-Score: .429**
**ROC AUC: .852**
**PR AUC: .421**

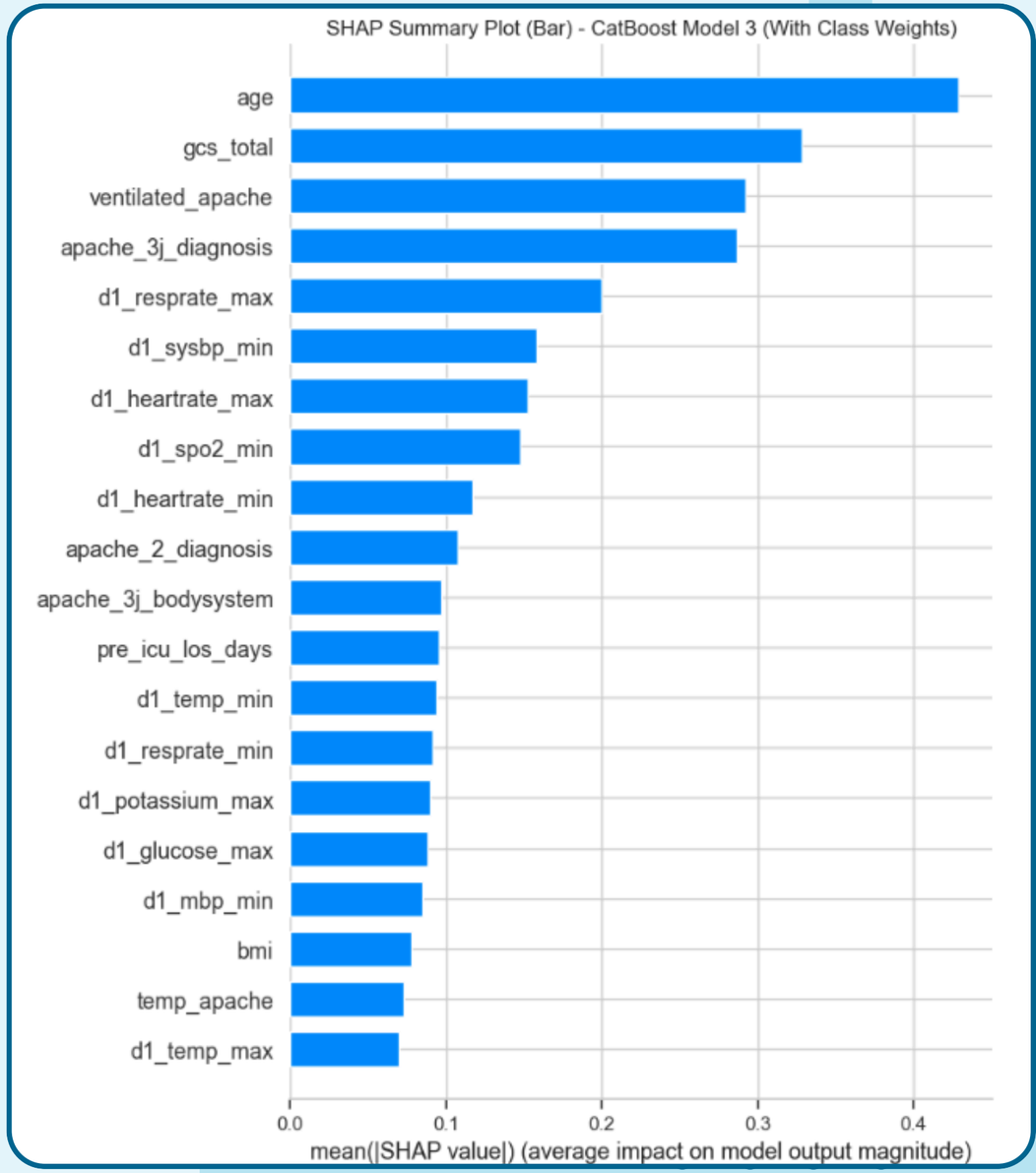
**Recall at 90% Threshold: .09**
**Precision: .163**
**F1-Score: .277**



# CATBoost: Class Weights

**Accuracy:** .812
**Recall:** .758
**Precision:** .276
**F1-Score:** .405
**ROC AUC:** .857
**PR AUC:** .455

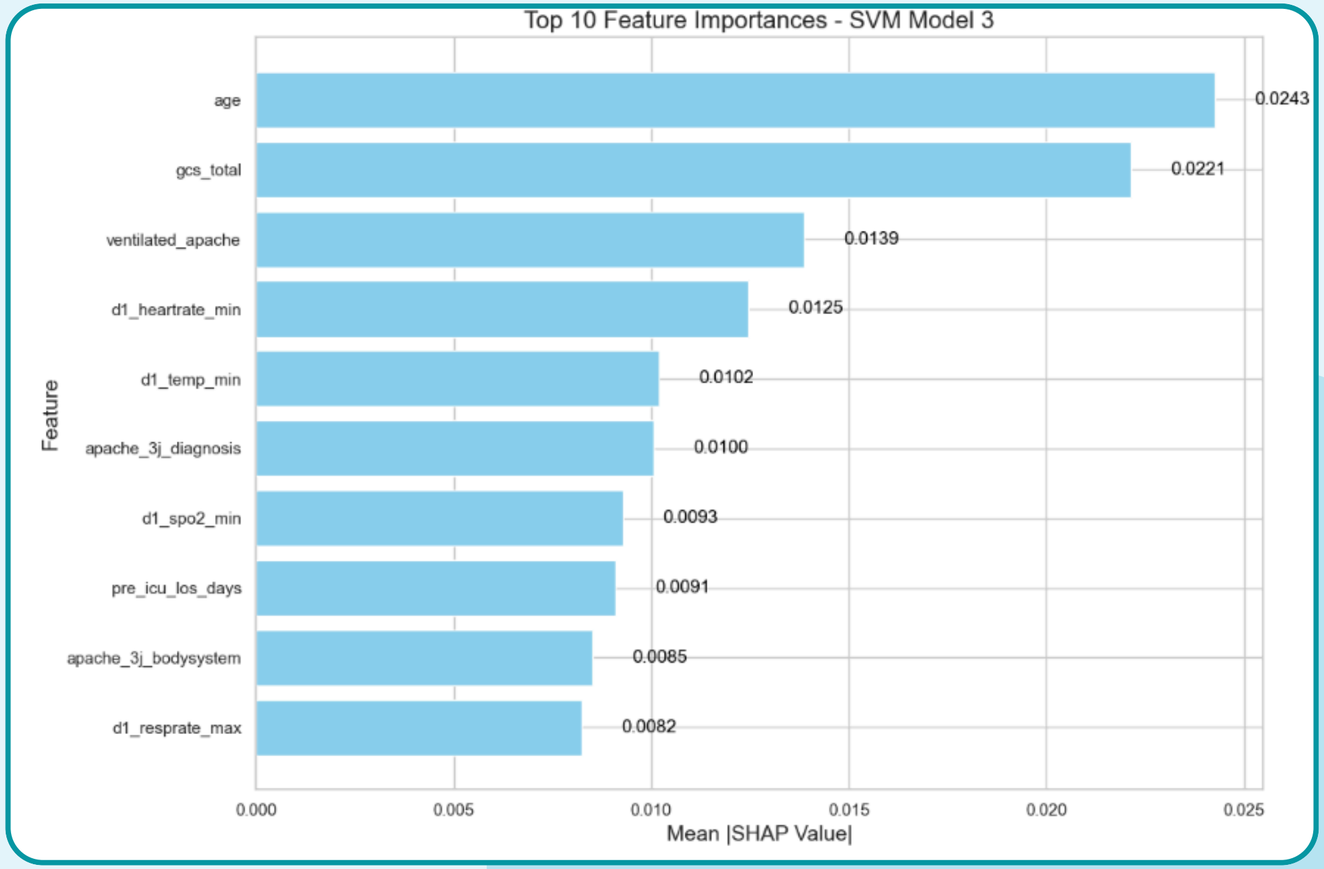
**Recall at 90% Threshold:** .29
**Precision:** .191
**F1-Score:** .315



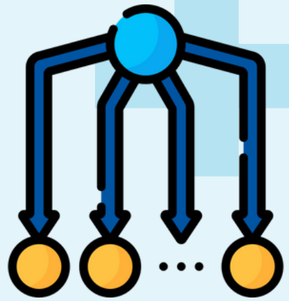
# Support Vector Machines: Class Weights

**Accuracy:** .838
**Recall:** .674
**Precision:** .297
**F1-Score:** .412
**ROC AUC:** .857
**PR AUC:** .368

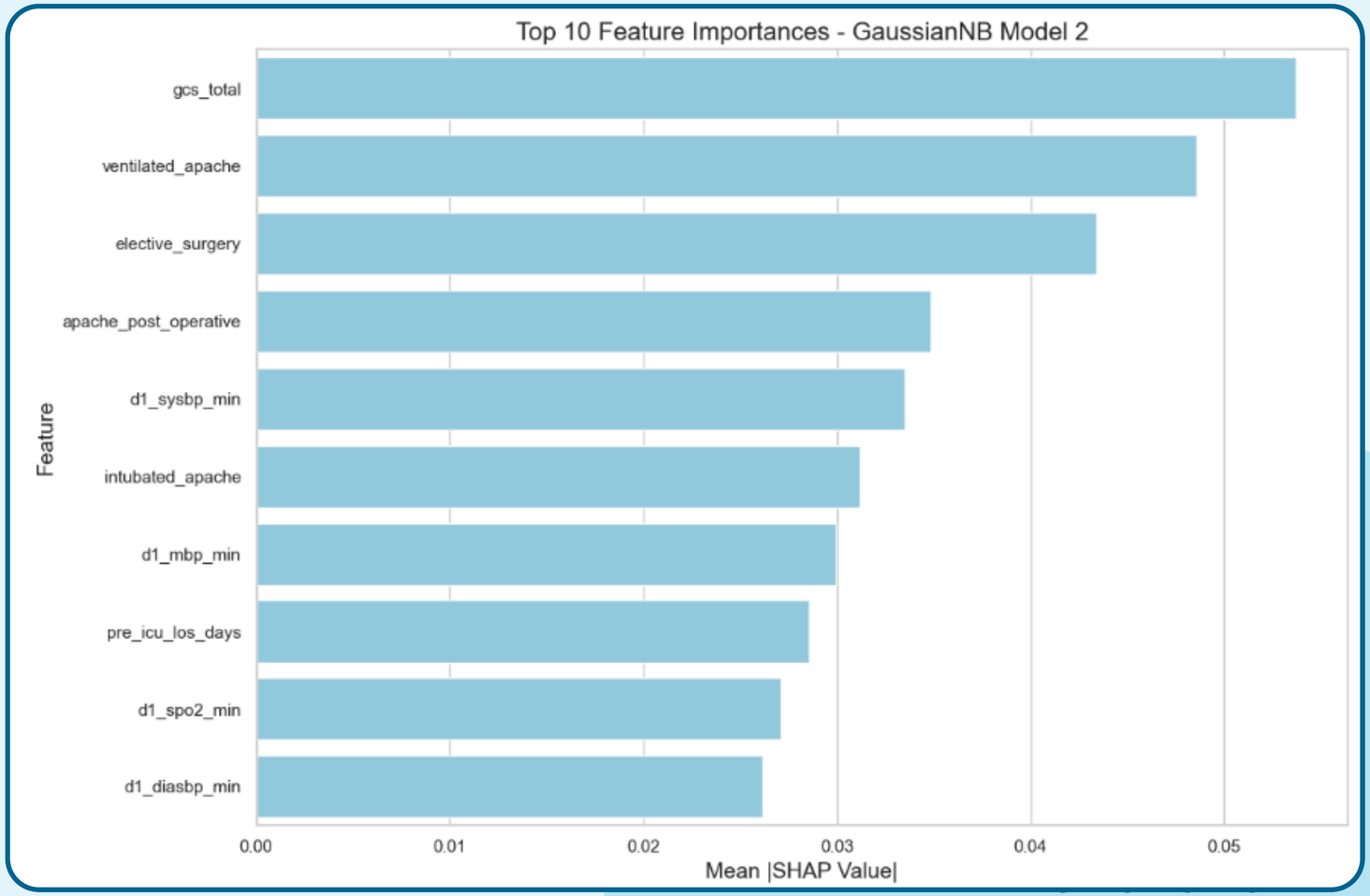
**Recall at 90%**
**Threshold:** .04
**Precision:** .170
**F1-Score:** .286



# Gaussian Naïve Bayes: SMOTENC

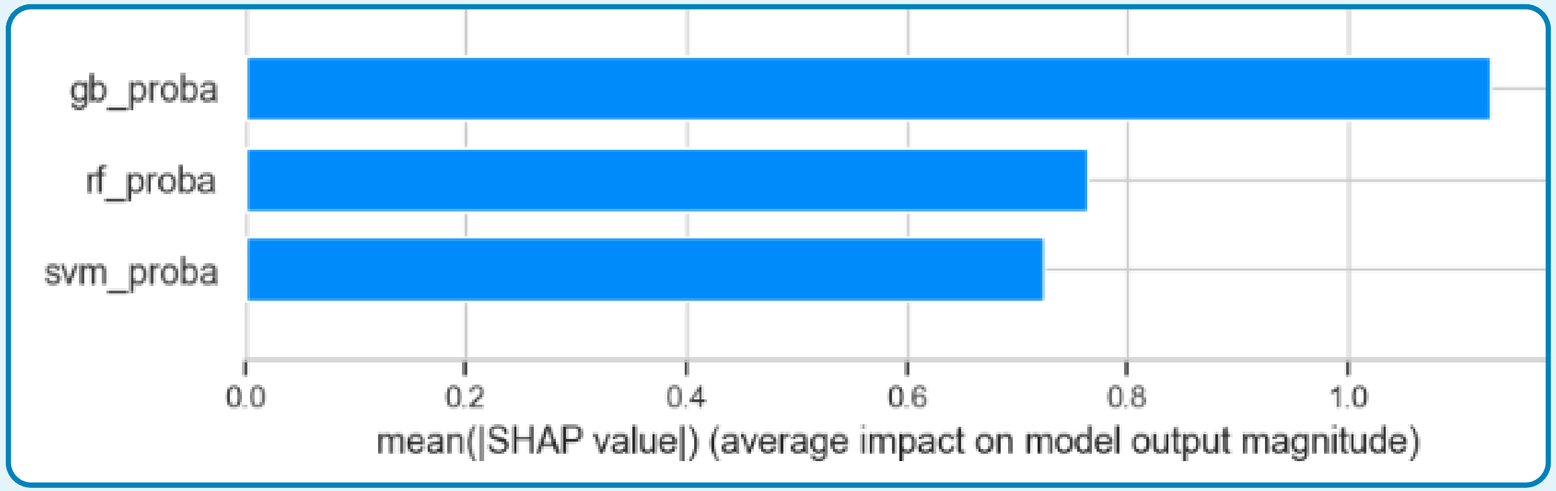


**Accuracy:** .771
**Recall:** .700
**Precision:** .224
**F1-Score:** .339
**ROC AUC:** .795
**PR AUC:** .326

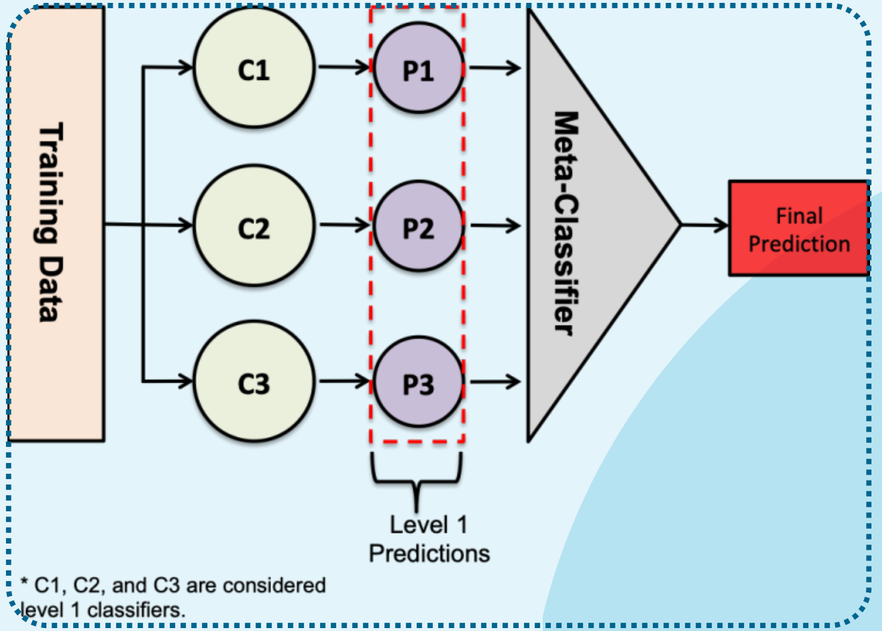


# Stacking Classifier: Class Weights

**Accuracy:** .956
**Recall:** .946
**Precision:** .965
**F1-Score:** .955
**ROC AUC:** .990
**PR AUC:** .992

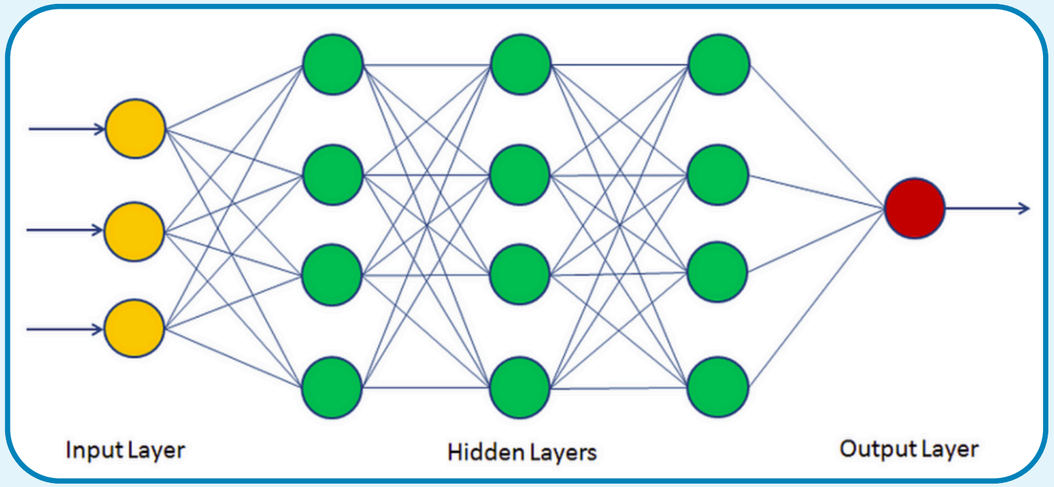


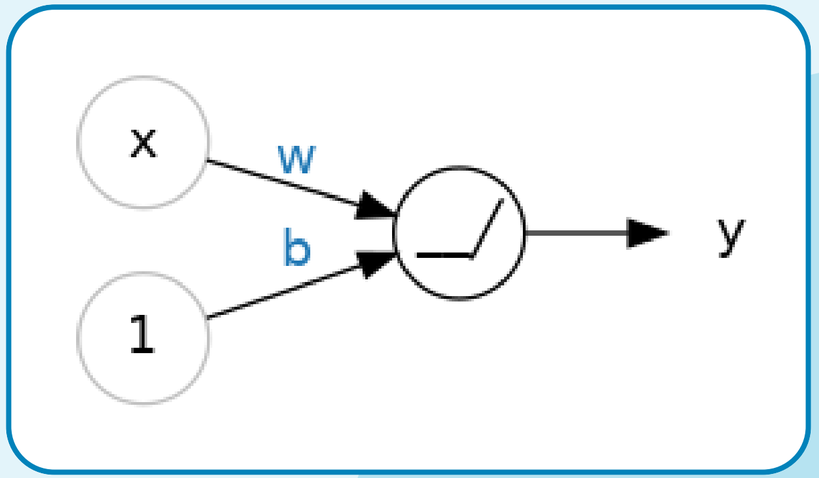

- Best metrics, but takes forever and not intuitive to interpret



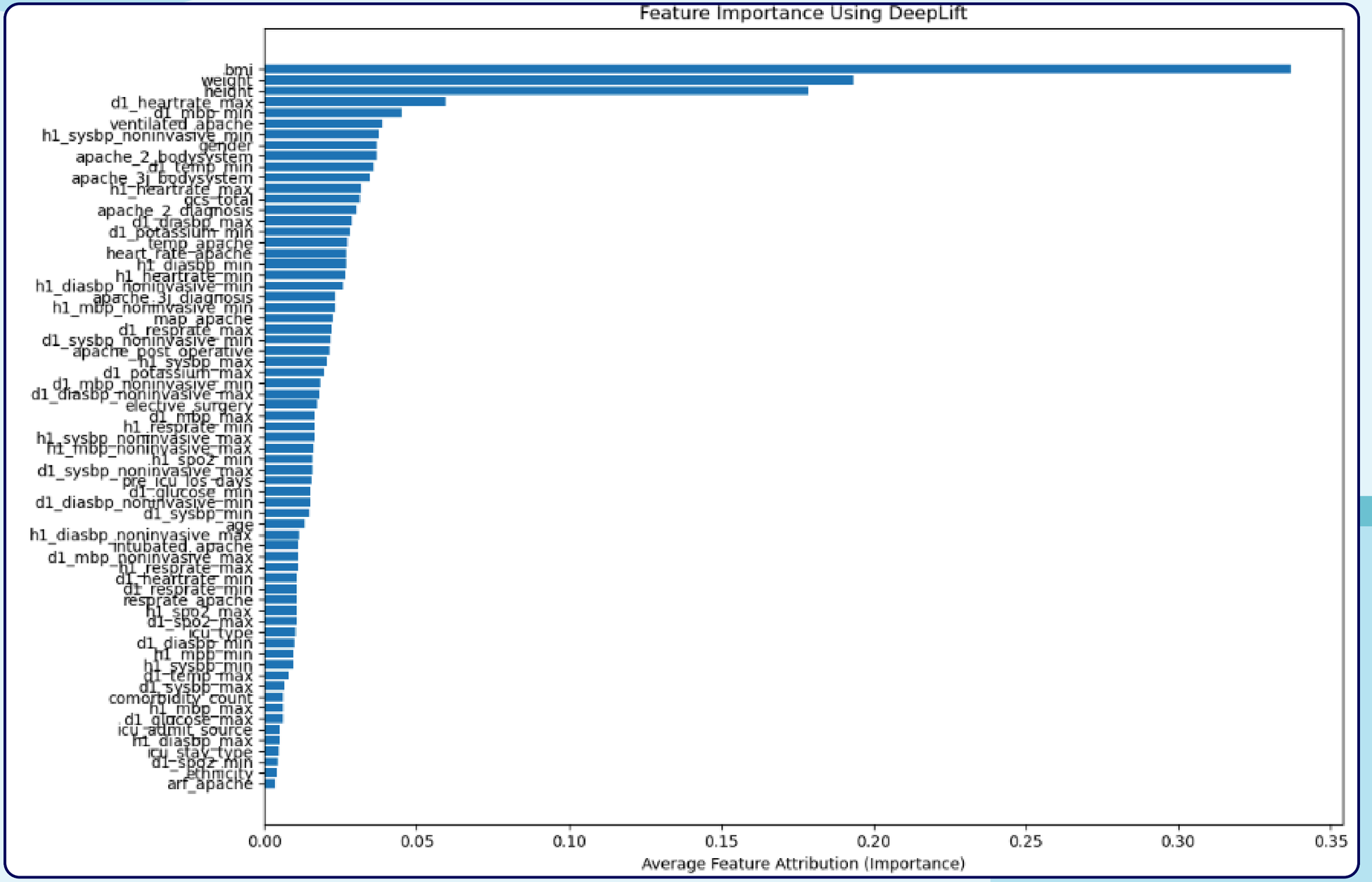
# MLPs

- Input → 128 → 64 → 1
- ReLU activation functions, sigmoid outer layer for binary classification
- Used early stopping and dropout after each hidden layer
- Tried Optuna to hypertune - took too long



# MLP - Feature Importance



# Best Model:
## MLP with SMOTENC

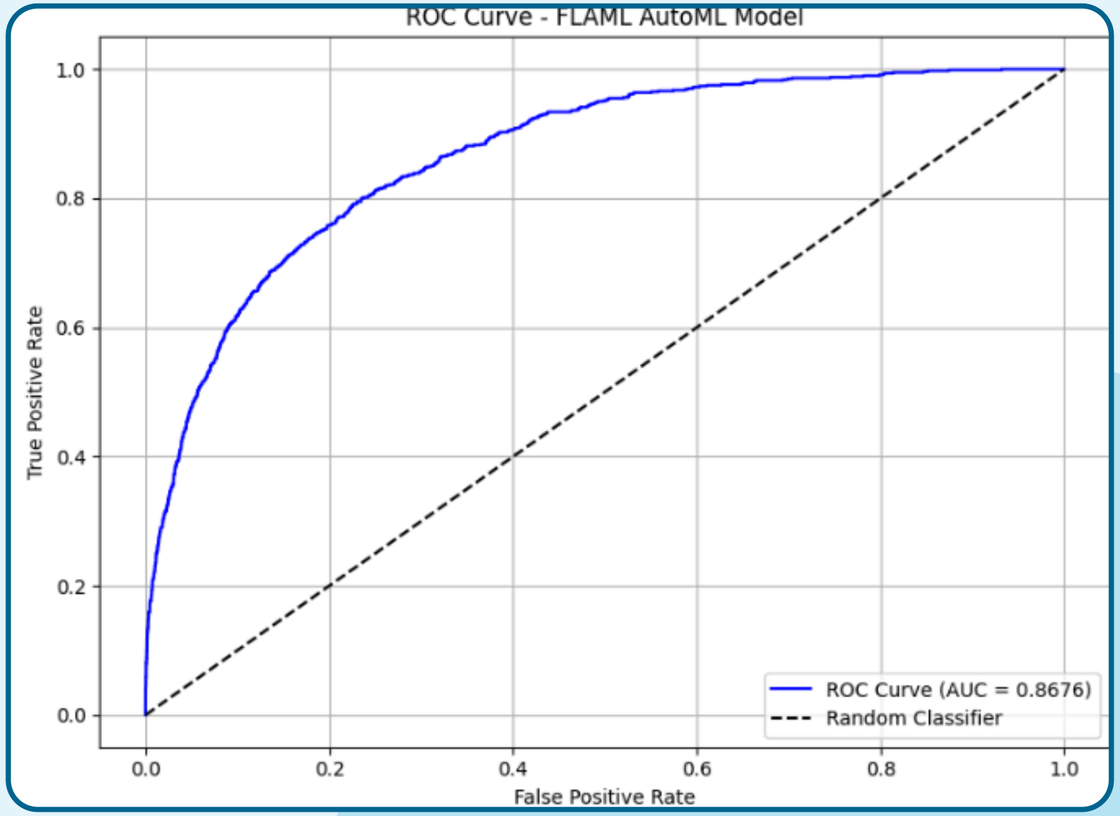
**Accuracy:** .9271
**Precision:** .9457
**Recall:** .9018
**F1-Score:** .92
**ROC AUC:** .985
**PR AUC:** .9837

**Threshold adjusting not required!**

# Extensions

# AutoML Using FLAML

- **Best Model:** XGBoost classifier
- 220 Trees with a max leaves of 128
- 65% of features used each tree
- **Learning Rate:** .12
- **Accuracy:** .92
- **Precision:** .61
- **Recall:** .30
- Overall - a flexible and complex tree



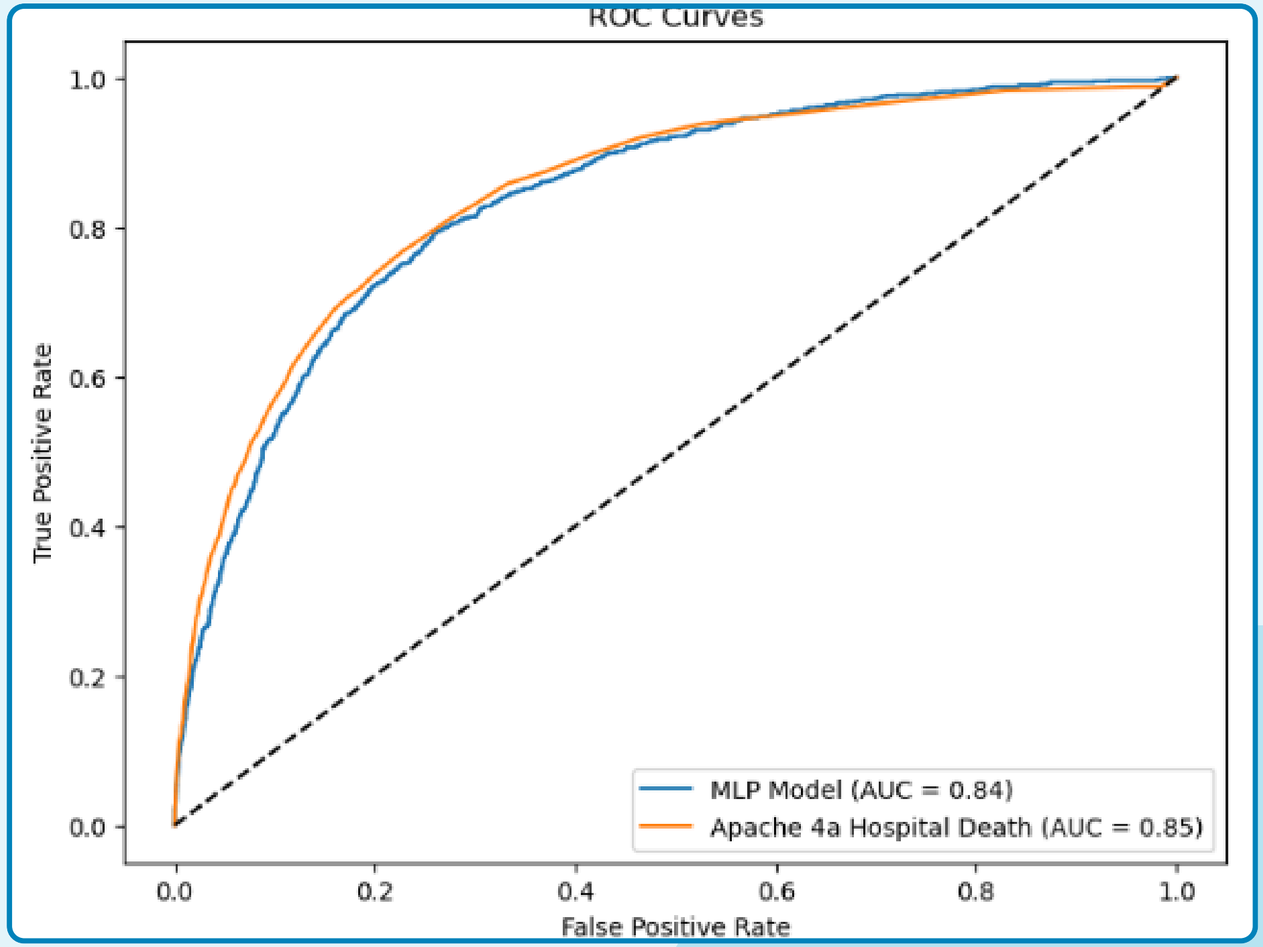
# Our Model vs APACHE

## MLP with SMOTENC

**Accuracy**: .86
**Precision**: .31
**Recall**: .58
**ROC AUC**: .8430

## APACHE

**Accuracy:** .90
**Precision:** .44
**Recall:** .41
**ROC AUC:** .8457



# Reinforcing Our Observations

Our best predictive models yield the following as the primary factors influencing ICU mortality :

***Age, GCS, BMI***.


**Can these conclusions be supported by existing research?**

- Leveraged publically available medical literature on PubMed
- Database contains **> 37 million** citations and abstracts.

# Reinforcing Our Observations

## The process was as follows :

- Retrieved abstracts for the **top 200 PubMed articles**
  - based on the query (e.g. '"ICU Mortality" AND ("Age" OR "Glasgow Coma Scale" OR "GCS" OR "BMI" OR "Height" OR "Weight")')
  - via the **PubMed API**.


- The **top 50 most relevant articles** were found
  - by calculating the cosine similarity score b/w each abstract and our query.
  - The query and abstracts were embedded using the **AllenAI SPECTER** model.

# Reinforcing Our Observations

- The **top 3 sentences** most relevant to the query were ranked and extracted from each abstract, using the **BM25** algorithm.

- The extracted top 3 sentences were passed to a **Hugging Face BART** model for **summarization**.

- The summaries, along with the extracted sentences and original PubMed IDs, were saved to a CSV file.


# Reinforcing Our Observations

**Cosine similarity scores output for the top 50 most relevant PubMed articles.**

| PubMed ID | Abstract | Similarity |
|---|---|---|
| 39482088 | 1. Am J Crit Care. 2024 Nov 1;33(6):446-454. doi: 10.4037/ajcc2024898. | 0.7626955509185791 |
| 38134715 | 1. J Crit Care. 2024 Apr;80:154509. doi: 10.1016/j.jcrc.2023.154509. Epub | 0.7614608407020569 |
| 37605448 | 1. AIDS. 2023 Nov 15;37(14):2169-2177. doi: | 0.7605999708175659 |
| 39567222 | 1. HIV Med. 2024 Nov 20. doi: 10.1111/hiv.13737. Online ahead of print. | 0.7553196549415588 |
| 39496882 | 1. Neurocrit Care. 2024 Nov 5. doi: 10.1007/s12028-024-02150-8. Online | 0.7510473132133484 |

**Some examples of summarized articles that back up our findings.**

| PubMed ID | Extracted Sentences | Summarized Text |
|---|---|---|
| 39266274 | BACKGROUND: Age is a significant consideration for intensive care unit (ICU) admission. CONCLUSIONS: Age had a significant impact on ICU mortality in our cohort of critically ill patients. The objective of the present study was to determine the impact of increasing age on ICU mortality. | Age is a significant consideration for intensive care unit (ICU) admission. Age had a significant impact on ICU mortality in our cohort of critically ill patients. The objective of the present study was to determine the impact of increasing age on ICu mortality. |
| 38662674 | Significant mortality-contributing factors included medical diagnosis, Glasgow Coma Scale score, sepsis/septic shock, sedation use, multiple-organ dysfunction syndrome, and cardiovascular disease. Ethiopia experiences a high intensive care unit (ICU) mortality rate. Results: The pooled mortality rate among adult ICU patients undergoing MV was 48.61% (95% CI: 40.82, 56.40%). | Ethiopia experiences a high intensive care unit (ICU) mortality rate. Significant mortality-contributing factors included medical diagnosis, Glasgow Coma Scale score, sepsis/septic shock, sedation use and multiple-organ dysfunction syndrome. |
| 38905370 | Sequential organ failure assessment (SOFA) scores and acute physiology and chronic health evaluation (APACHE) II were used to assess disease severity. The HALP score was not associated with ICU LOS or a significant prognostic factor for mortality. APACHE-II, SOFA, and mNUTRIC were the strongest prognostic indices for ICU mortality, with mNUTRIC having the highest sensitivity and negative predictive value. | The HALP score was not associated with ICU LOS or a significant prognostic factor for mortality. APACHE-II, SOFA, and mNUTRIC were the strongest prognostic indices for ICU mortality. mN UTRIC had the highest sensitivity and negative predictive value. |

# 05

# Lessons/ Conclusions

# What Model is Best at Predicting Hospital Deaths?

## MLP with SMOTENC

**Accuracy: .9271**
**Precision: .9457**
**Recall: .9018**
**F1-Score: .92**
**ROC AUC: .985**
**PR AUC: .9837**

- For all models we can **adjust threshold to balance precision and recall**
  - results in Low False Negatives at expense of more False Positives
- **Most important features:** Age, ventilated, Glasgow Coma Scale Total

# Lessons Learned

- **Preprocessing is hard and takes time:**
  - **feature engineering helps combat the curse of dimensionality**
- **Addressing class imbalance is crucial for improving model performance on minority classes**
- **Balancing recall and precision requires careful consideration to avoid overwhelming resources with false positives**

# Future Work

- Look at feature interactions and other ways to **perform dimensionality reduction to capture complex relationships**
- Utilize techniques like **Grid Search or random search to fine-tune model** parameters for optimal performance
- **Conduct fairness assessments** to ensure the model does not inadvertently introduce biases, especially after collapsing ethnic categories

# Thank You!

Questions?