HEART FAILURE MORTALITY PREDICTION

DSI23 - Capstone

Tay Yi Li

CONTENTS

1

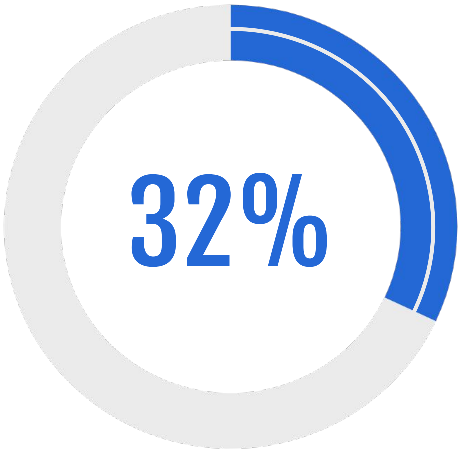
Background

Overview: Heart Failure

- Heart muscle does not pump blood well
- No cure
- **Cardiovascular diseases (CVDs)** frequently ends in **Heart Failure**

17,900,000

Deaths by CVDs in 2019



CVDs accounted for 32% of all deaths in 2019

CVDs can be prevented/controlled if we..

Address Behavioural Risk Factors

- Unhealthy diet
- Lack of exercise
- Smoking, etc..

Manage Underlying Conditions

- High Blood Pressure
- Diabetes, etc..

Since Heart Failure is commonly caused by CVDs

Controlling and Managing **CVDs'** Risk Factors to prevent deaths

→

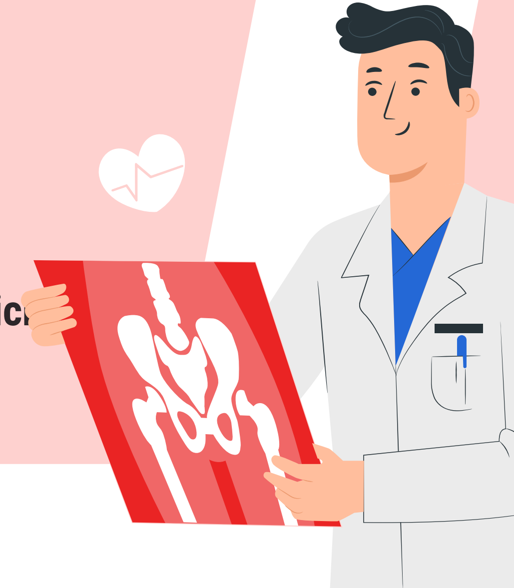
Controls and prevents death by **Heart Failure**

The key is early detection and management!

Problem Statement

With earlier care and attention given, mortality by heart failure can be prevented. The Department of Cardiology tasked the newly established Data Science Department to find a way to identify patients with high risks of mortality by Heart Failure through use of data science to enable them to provide necessary preventive care and attention for the patients early.

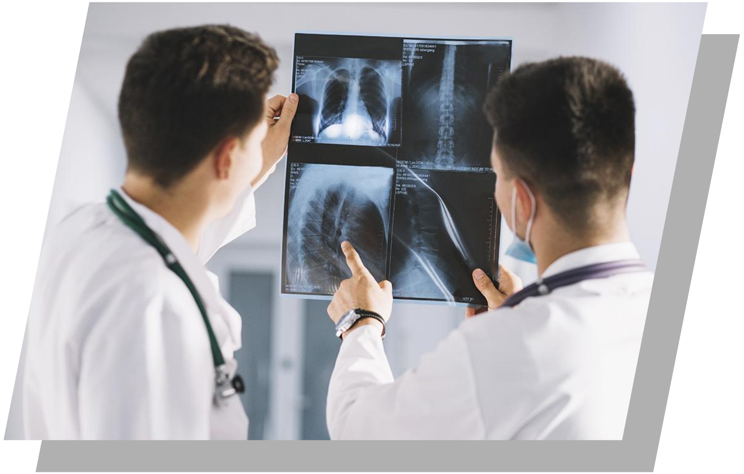
To achieve this, the project aims to build a classifier which uses patients' health conditions to accurately predict mortality by Heart Failure.



Objectives

Build a classifier using patients' health conditions to accurately predict mortality by Heart Failure.

Model will help identify patients most in need of earlier care and attention.



Metrics used:

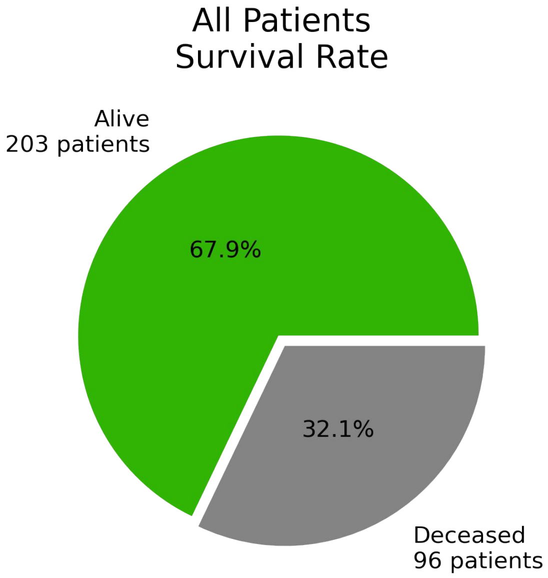
1. F1 score
2. Precision-**Recall** score
3. Train/Test Accuracy

2

EDA

Overview – Dataset

- Dataset obtained from Kaggle
- 13 features in total including target variable
- No clean up required
- Imbalanced dataset



Overview – Target Variable

Target variable: death_event

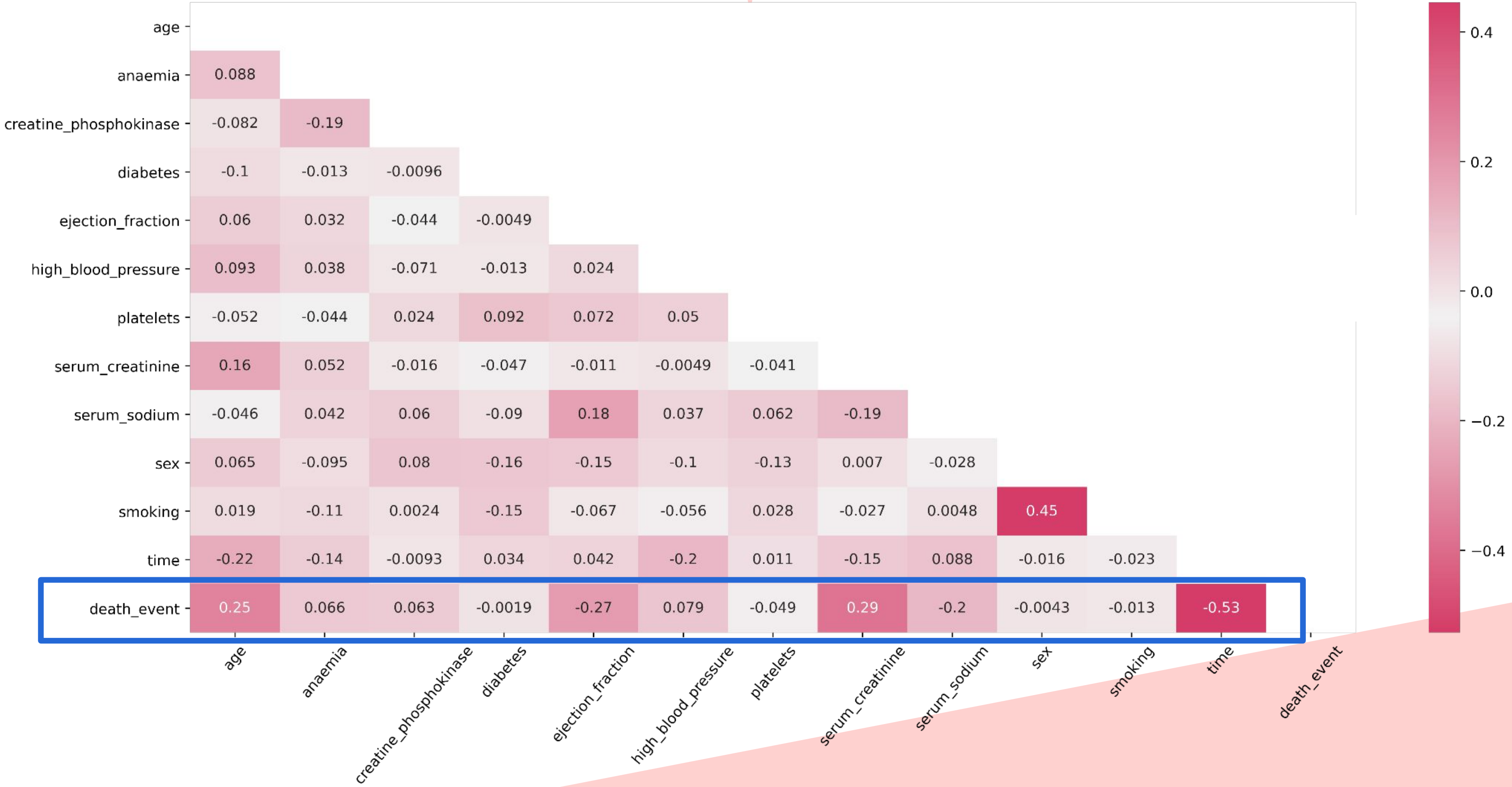
death_event is binary and indicates whether the patient survived

death_event		
Value in Dataset	**Heart Failure Survival**	**Class**
1	Deceased	Positive
0	Alive	Negative

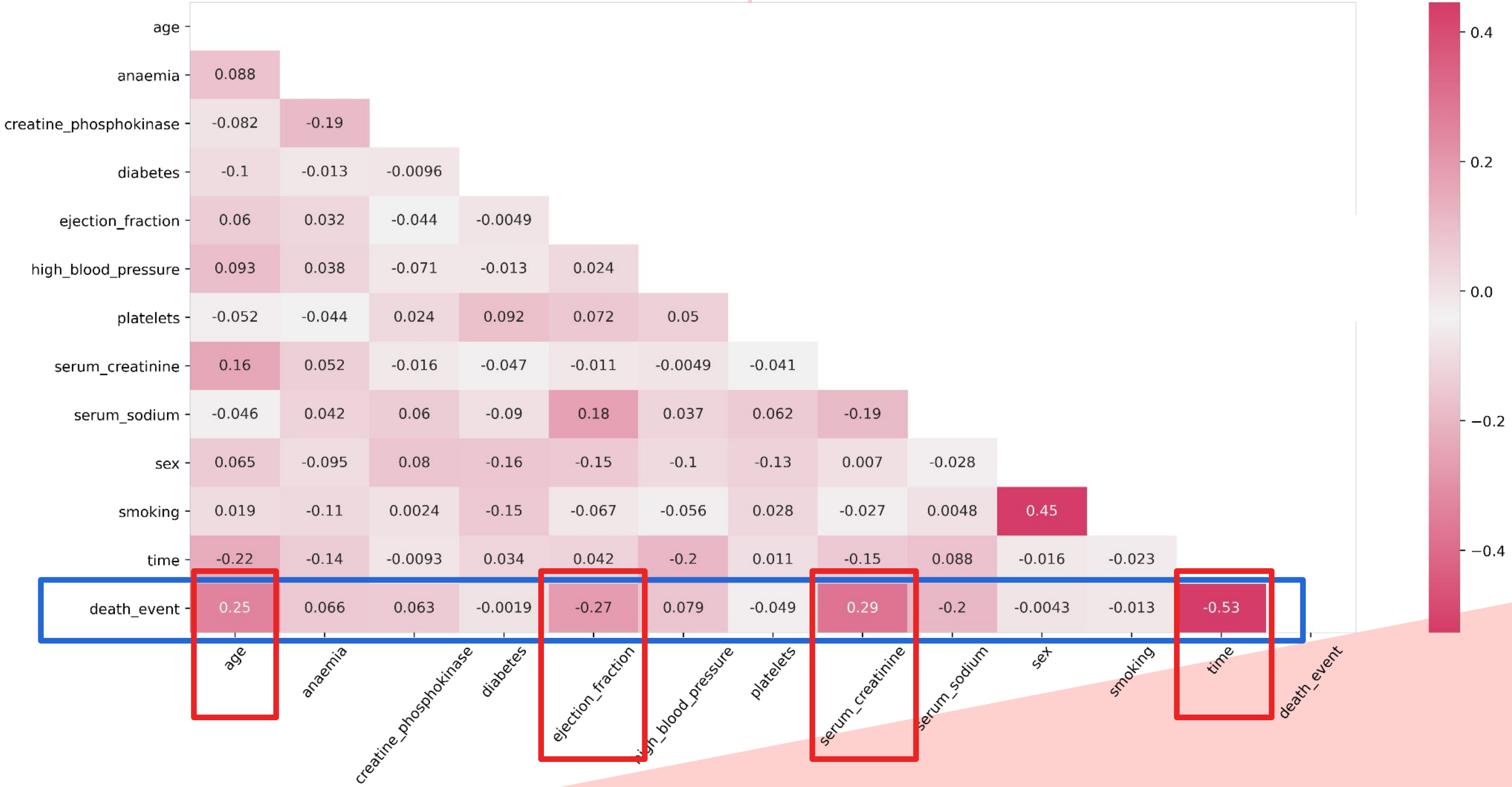
Overview – Research

Feature Name	Description Summary	Relationship with Heart Failure
Anaemia	Lower than usual red blood cells/hemoglobin	Untreated anaemia potentially leads to heart failure
Creatine Phosphokinase (CPK-MB)	Enzyme that leaks into blood when heart is damaged	Elevated levels indicate injury to heart
Diabetes	Inability to regulate blood sugar	Increases chances of getting CVDs
Ejection Fraction (EF)	Percentage of blood pumped out of heart	Low EF indicates heart is not working well, heart failure is likely occurring
High Blood Pressure (HBP)	Blood pressure is consistently higher than normal	HBP increases risk of CVDs and heart attack
Platelets	Small blood cells that forms clots to stop bleeding	Too many platelets may result in heart attack
Serum Creatinine	Waste product filtered out of blood by kidneys	Elevated levels may be indicator of heart failure
Serum Sodium	Amount of sodium in blood	Low levels may be indicator of heart failure

Feature Selection



Feature Selection



3

Modelling

Metrics

1. F1 Score
2. Precision-**Recall** Score
3. Train/Test Accuracy

F1 Score as main evaluation metric

$$F1score = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Indication of model's accuracy on the dataset and overall performance
- Weighted average of Precision-Recall score
- Affected by Recall score

Good model = High F1 score

Precision-Recall Score with focus on Recall

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- Trade-off between Precision and Recall scores
- Between False Positive and False Negative, we want to have low False Negative
- False Negative = predict to survive, but passed away
- False Positive = predict to pass away, but survived
- False Negative patients are not identified as “high risk of mortality by Heart Failure”
- Patient does not receive earlier care and attention that may save their life

False Negatives as low as possible = Recall score as high as possible

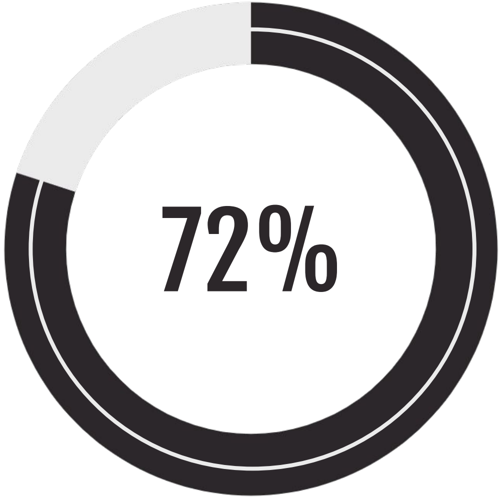
Accuracy score not used as main evaluation metric

$$Accuracy = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}$$

- Overview of model's ability in predicting majority and minority class combined
- Unreliable scoring for imbalanced datasets
- Majority class would have high number of correct predictions, leading to high Accuracy score
- Masks model's inability to correctly classify minority class
- Not really affected by Recall score

Train/Test Accuracy to be compared to check for underfitting/overfitting

Baseline Model - Logistic Regression

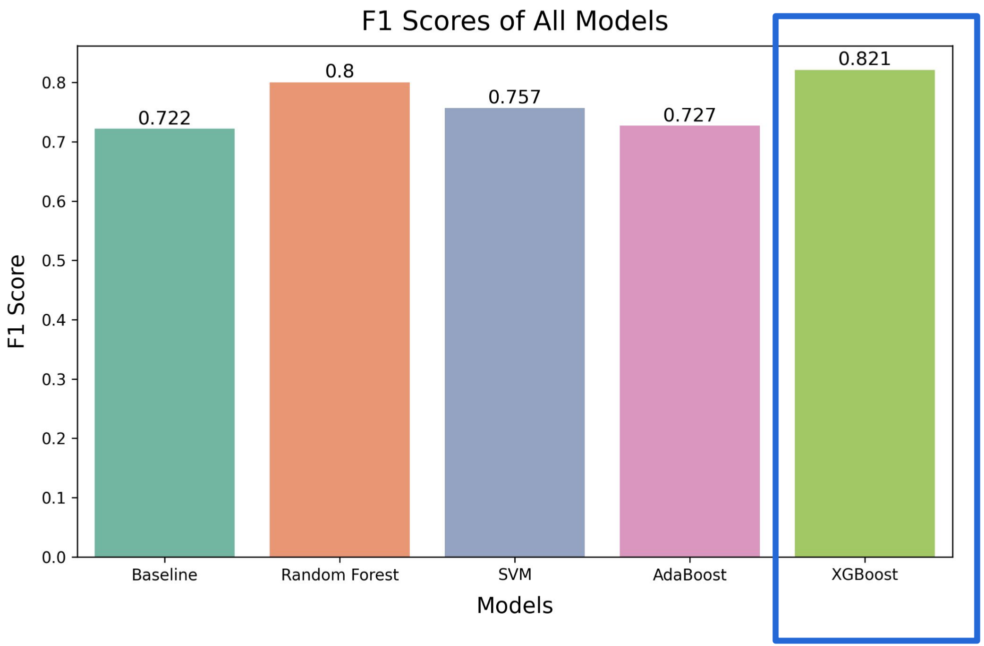


F1 Score of 72%

Models

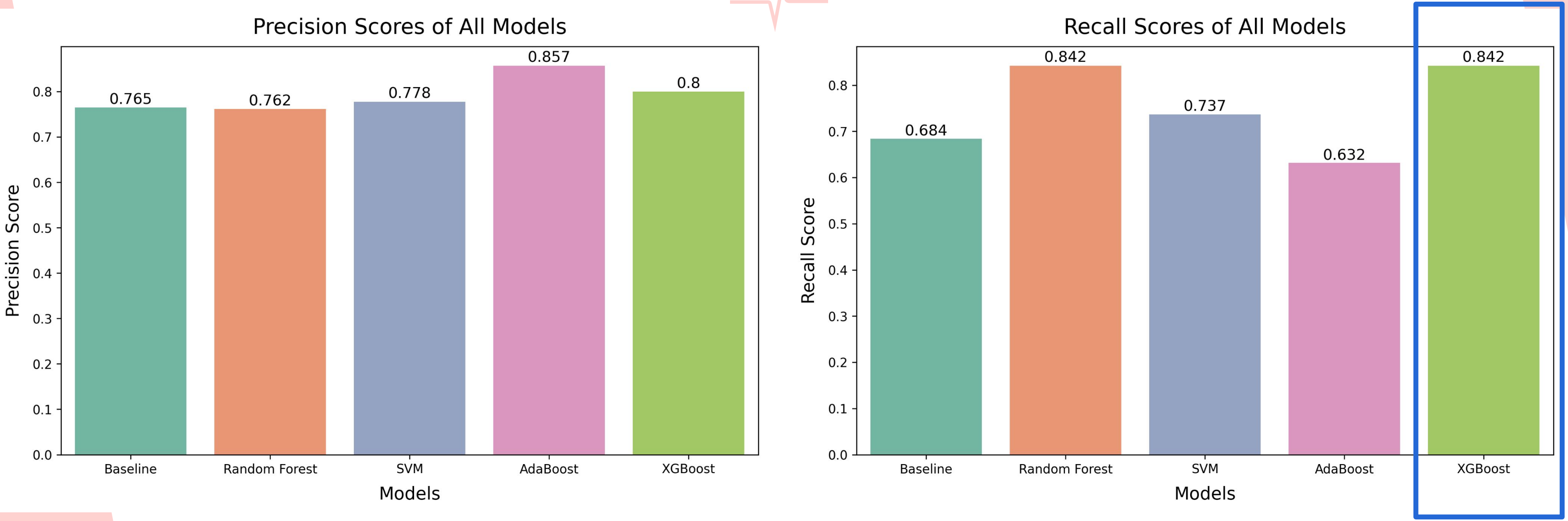
1. Random Forest
2. Support Vector Classifier
3. AdaBoost
4. XGBoost

Models Comparison – F1 Score



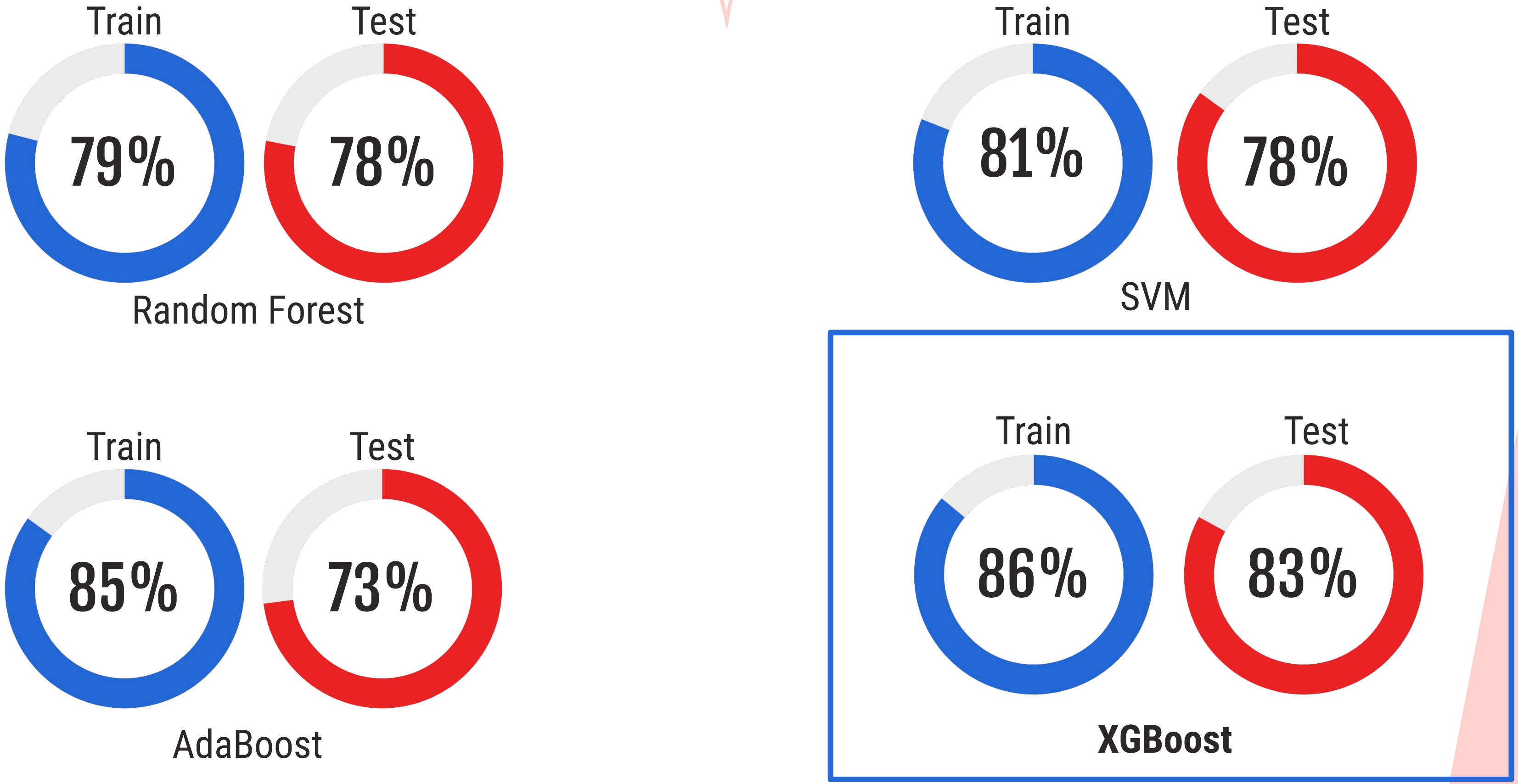
XGBoost has highest F1 score of 0.821

Models Comparison - Precision-**Recall** Score



XGBoost has highest **Recall** score of 0.842

Models Comparison – Accuracy Scores



Models Comparison – Summary

	F1 score	Precision	Recall	Train Accuracy	Test Accuracy
Baseline	0.722	0.765	0.684	0.820	0.833
Random Forest	0.800	0.762	0.842	0.795	0.783
SVM	0.757	0.778	0.737	0.816	0.850
AdaBoost	0.727	0.857	0.631	0.849	0.733
XGBoost	**0.821**	0.800	**0.842**	0.862	0.833

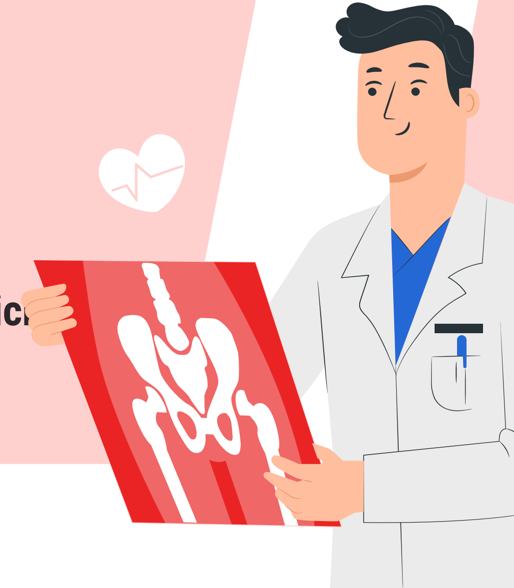
4

Conclusion

Problem Statement

With earlier care and attention given, mortality by heart failure can be prevented. The Department of Cardiology tasked the newly established Data Science Department to find a way to identify patients with high risks of mortality by Heart Failure through use of data science to enable them to provide necessary preventive care and attention for the patients early.

To achieve this, the project aims to build a classifier which uses patients' health conditions to accurately predict mortality by Heart Failure.



Conclusion - Model Uses

Best Performing: **XGBoost model**

- F1 score of 0.821
- Recall score of 0.842

Model may help with:

- **Faster identification of patients at highest risk of mortality from heart failure**
- **Allow more efficient allocation of appropriate attention and resources to patients who needs it most**

Conclusion – Recommendations

Recommendations for further improvement

- Tuning hyperparameters
- More rows of data
- More specific details on underlying conditions

Future Steps

- Modify model to generate likelihood of mortality from heart failure
- Apply model to other types of causes of death, like stroke

THANKS!

Any questions?