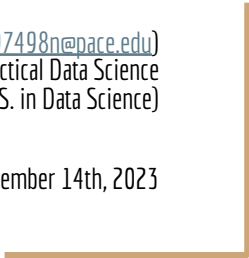


# Heart Failure Prediction - A pilot study project for Kaiser Permanente

Darsh Joshi ([dj97498n@pace.edu](mailto:dj97498n@pace.edu))  
CS 672 - Practical Data Science  
(M. S. in Data Science)

November 14th, 2023



# Agenda

- Executive Summary
  - Project Plan Recap
  - Data
  - EDA
-

# Executive Summary

In an effort to enhance patient care and preemptively identify at-risk individuals, the team has developed a sophisticated predictive model for heart failure, tailored specifically for the patient demographics of the Kaiser Permanente hospital chain. Utilizing state-of-the-art machine learning techniques and the rich medical datasets available, our model aims to provide clinicians with a powerful tool to assess heart failure risks, enabling timely interventions and optimizing resource allocation, ultimately driving better patient outcomes and reducing hospital readmissions.

---

# Project Plan Recap

Deliverable	Details	Due Date	Status
Data & EDA	Initial steps of the project: gathering data, cleaning, and visualising the data.	10/31/23	Completed
Methods, Findings, and Recommendations	Finding insights from the data and implementing model to define some initial recommendation.	11/14/23	Completed
Final presentation	Complete entire project and include all major findings in the presentation.	12/5/23	In Progress

Data



# Data Summary

- Data Source: Hospital Clinical Archive ( [Kaggle](#) )
  - Sample Size: 299 People (The sample size is relatively lower because it has to be accurate and medical issue is rare compare to others)
  - All available records from the source have been considered. (1st quarter 2023)
-

# Data Assumptions & more

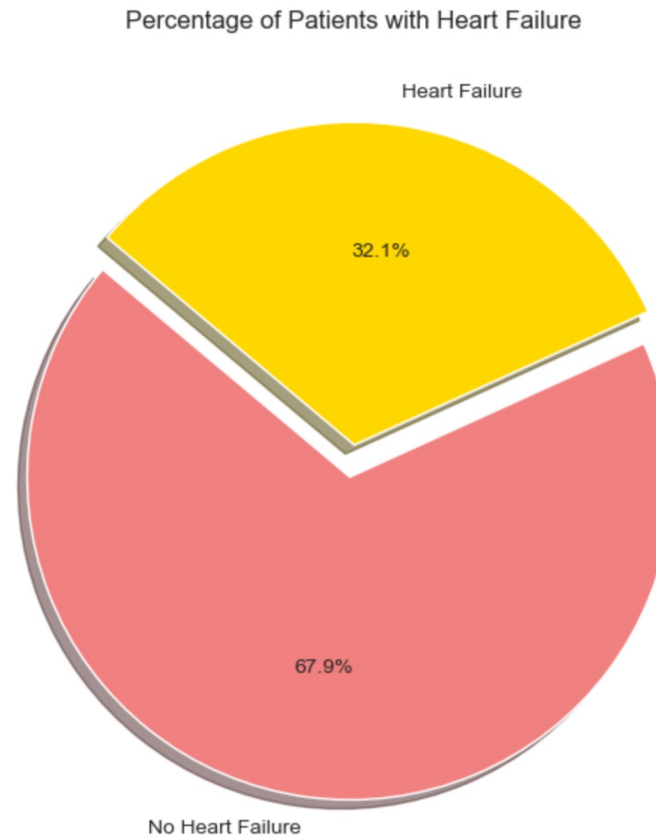
- Given that this data is sourced from Kaggle, we assume it has been anonymized and does not contain any personally identifiable information.
  - For the purpose of our analysis, we assume that the time column represents days. However, this should be verified with the original data provider.
- 
- For detailed information on each data columns refer to ([Appendix](#))
-

# Exploratory Data Analysis

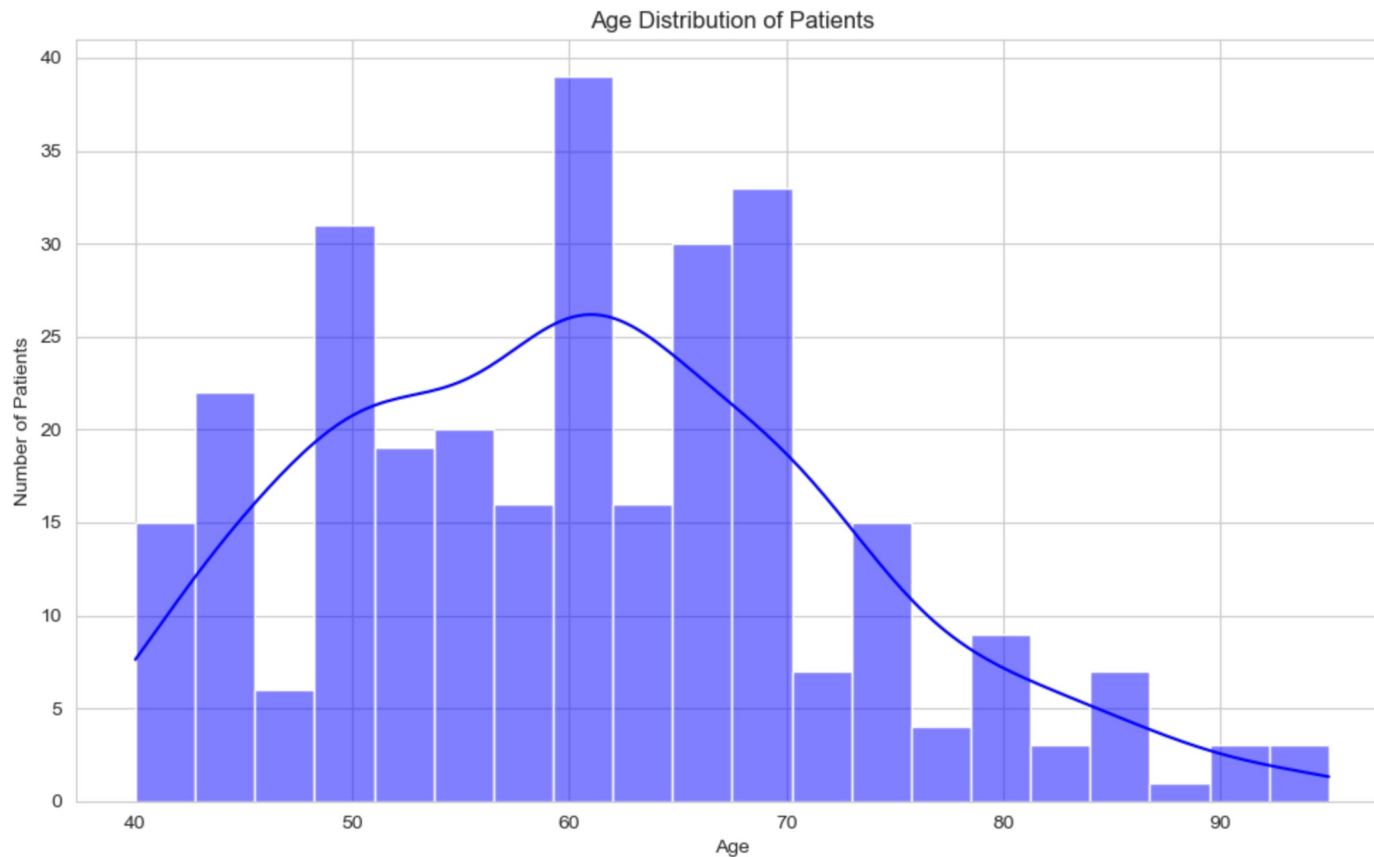


# Do we have a vast data to begin with?

This Pie chart represents the distribution of two types of data that we need in order to begin this analysis.



# Age Distribution of the people in dataset



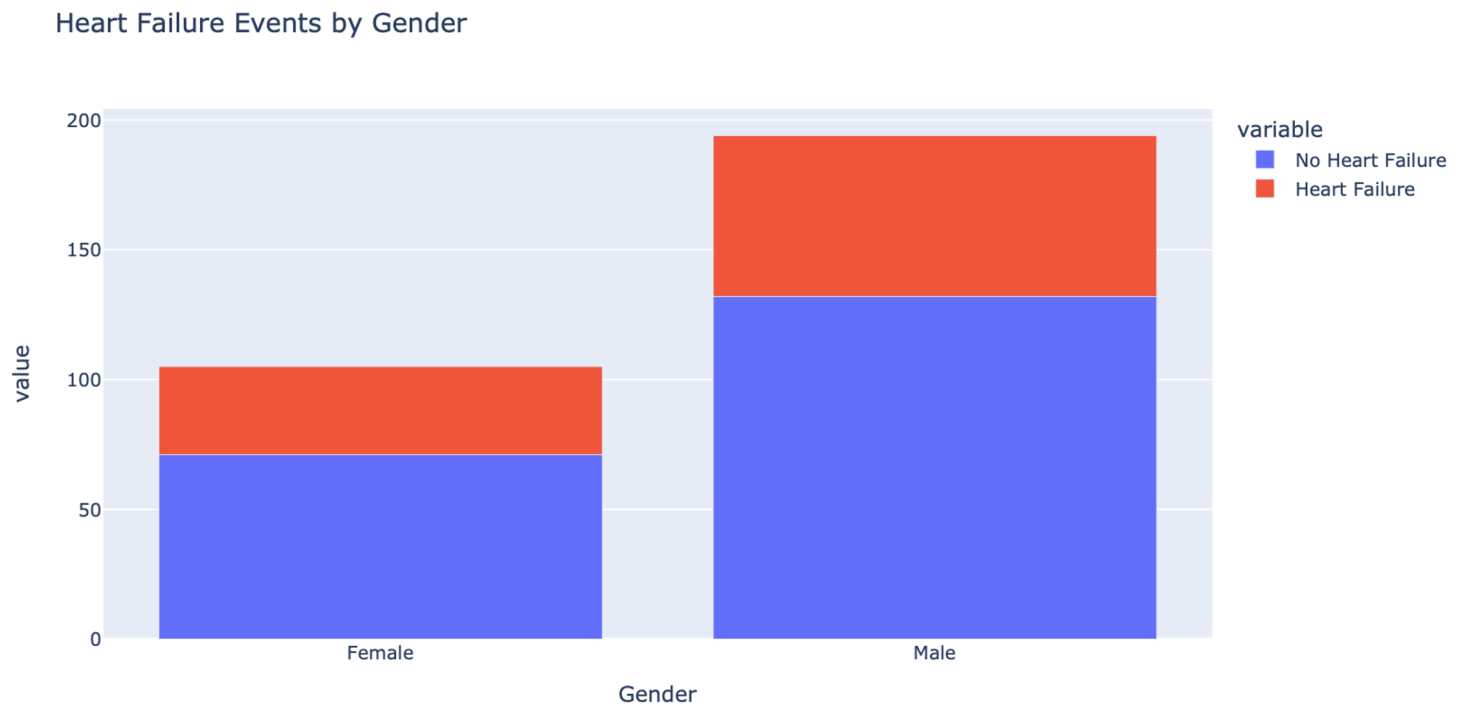
This chart shows what type of data we have about the age of patients.

It gives us crucial information regarding range of patients which might have heart related issues.

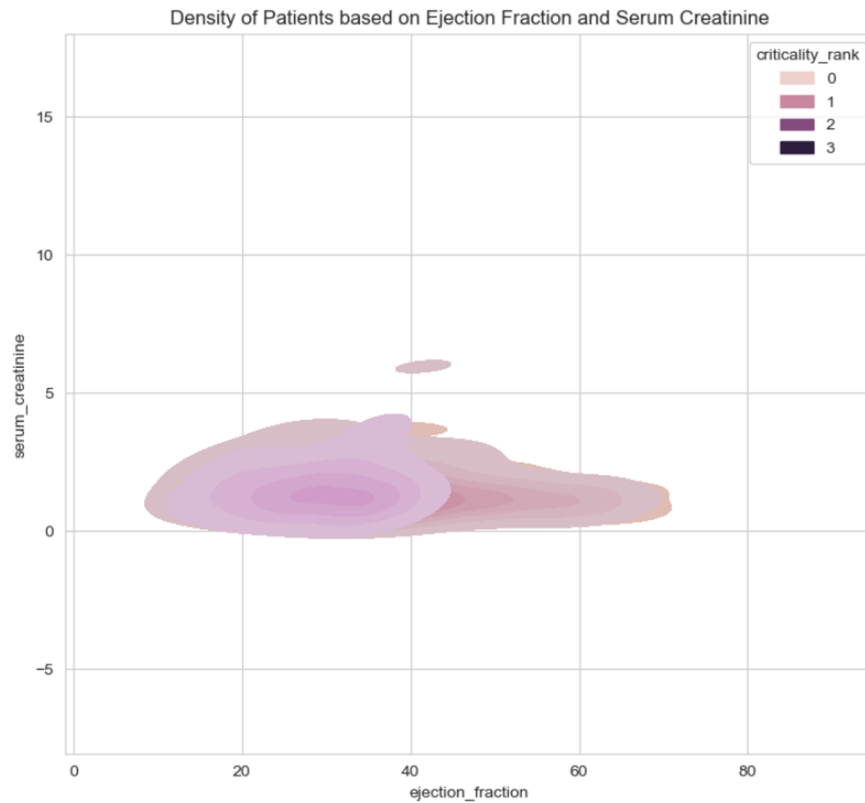
# Which Gender is more prone to heart failure?

This viz represents information regarding heart failures in geneders.

This information helps us regarding manual care for patients which higher failure ratio.



# How top features are aligned with heart failure?



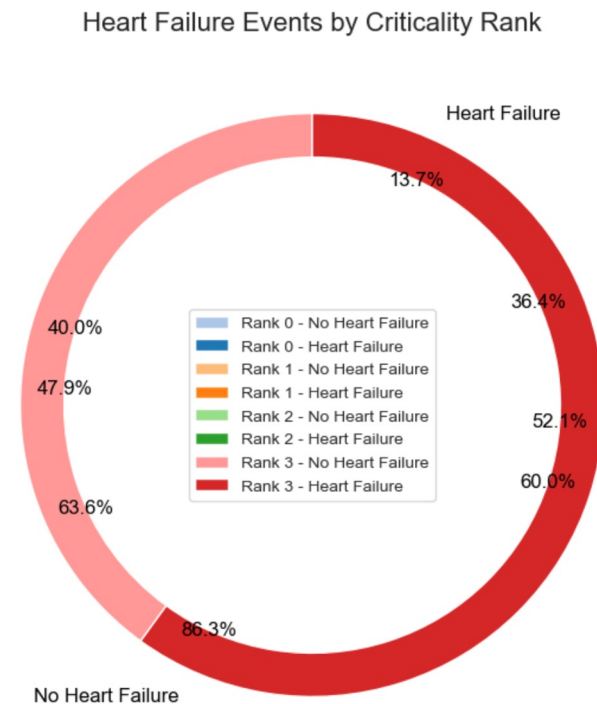
This data viz gives us an idea of our process if it's making any sense.

Closer this graphs is higher the sense it makes for us to use this data for ML models

# How important is Criticality ranking?

This plot is complex but represents very crucial information.

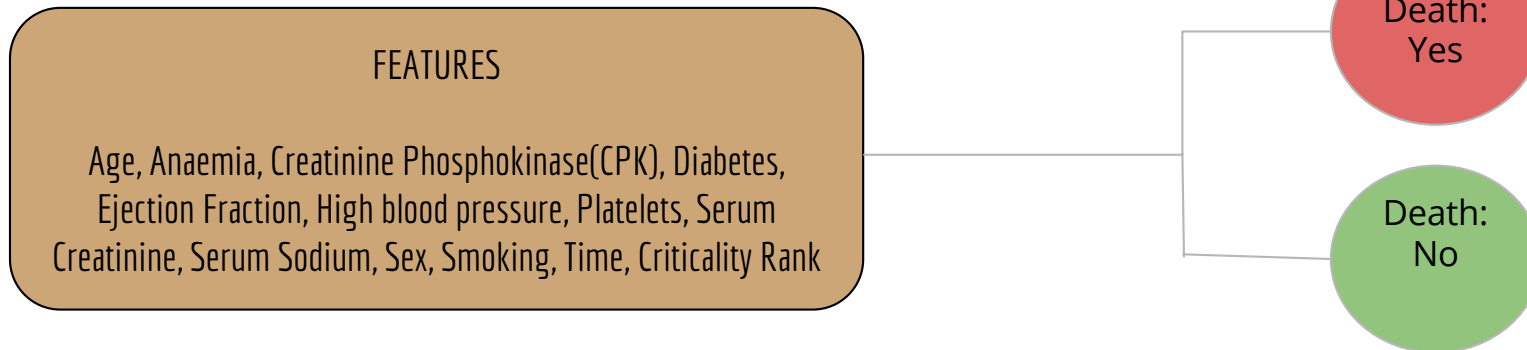
Rank 3 is the highest and only event that shows up here which means that it has higher correlation with Heart Failure. This is helpful for ML models



# Machine Learning Modeling

# What are we doing with this data?

- Now that we have enough and initial insights, we'll use it to create a Machine Learning Model using available data.
- What are we predicting exactly?



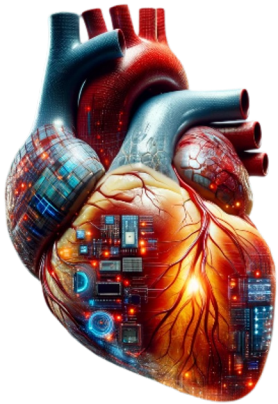
# Important Features and what are they?

- **Criticality Rank:** A new feature curated by our Team.
    - This feature is related to all the reported features and their critical values. I.e. If 2 of the reports are above or below critical range value is 2.
  - **Medical Report based features:** Anaemia, Creatinine Phosphokinase(CPK), Diabetes, Ejection Fraction, High blood pressure, Platelets, Serum Creatinine, Serum Sodium
  - **General Information based features:** Age, Sex, Smoking, Time
-



# Understanding Our Heart Failure Prediction Model

Think of our model as a high-tech heart guardian. It's like having a doctor who uses a supercomputer to watch over your heart health.



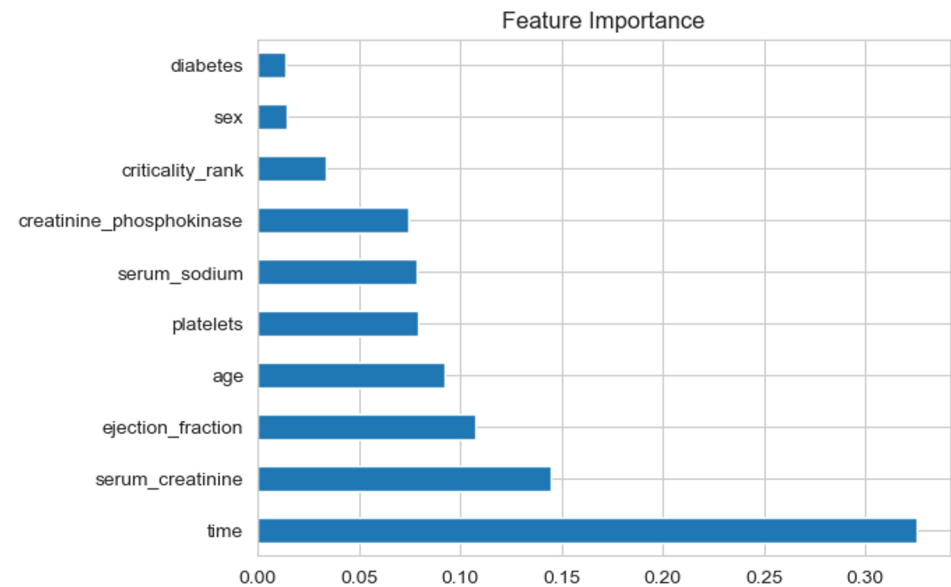
- **Data-Driven:** "Learns from many patients' health data."
  - **Factors Considered:** "Looks at age, health conditions, blood tests."
  - **Friendly and Reliable:** "95% accurate in identifying heart health risks."
  - **Model Name & Version :** "[Random Forest Baseline Model](#)."  
(Click on model for tech insights of model)
-

# Findings

# What that ML Model gives us?

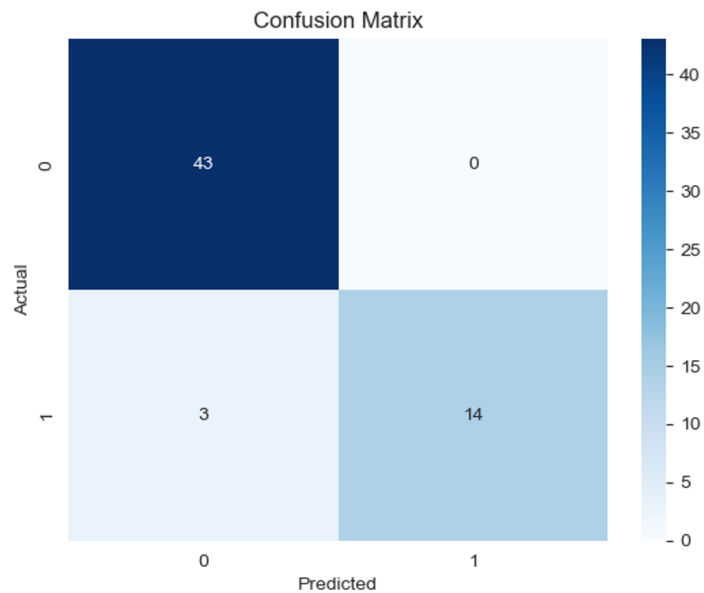
Identifying these key factors helps us focus on what's most important in managing and preventing heart-related issues in our patients.

- This model gave us importance of every aspects of tests for heart failure from most to least important.
- As per the model Time : Duration of Follow-up is the most important aspect of heart failure.
- Diabetes and sex although showing the least importance in the chart, they're important as well. The comparison is relative.



Note: The newly feature created by our data scientist is comparatively important then couple of pre-existing features.

# How Well Does Our Model Predict Heart Health?



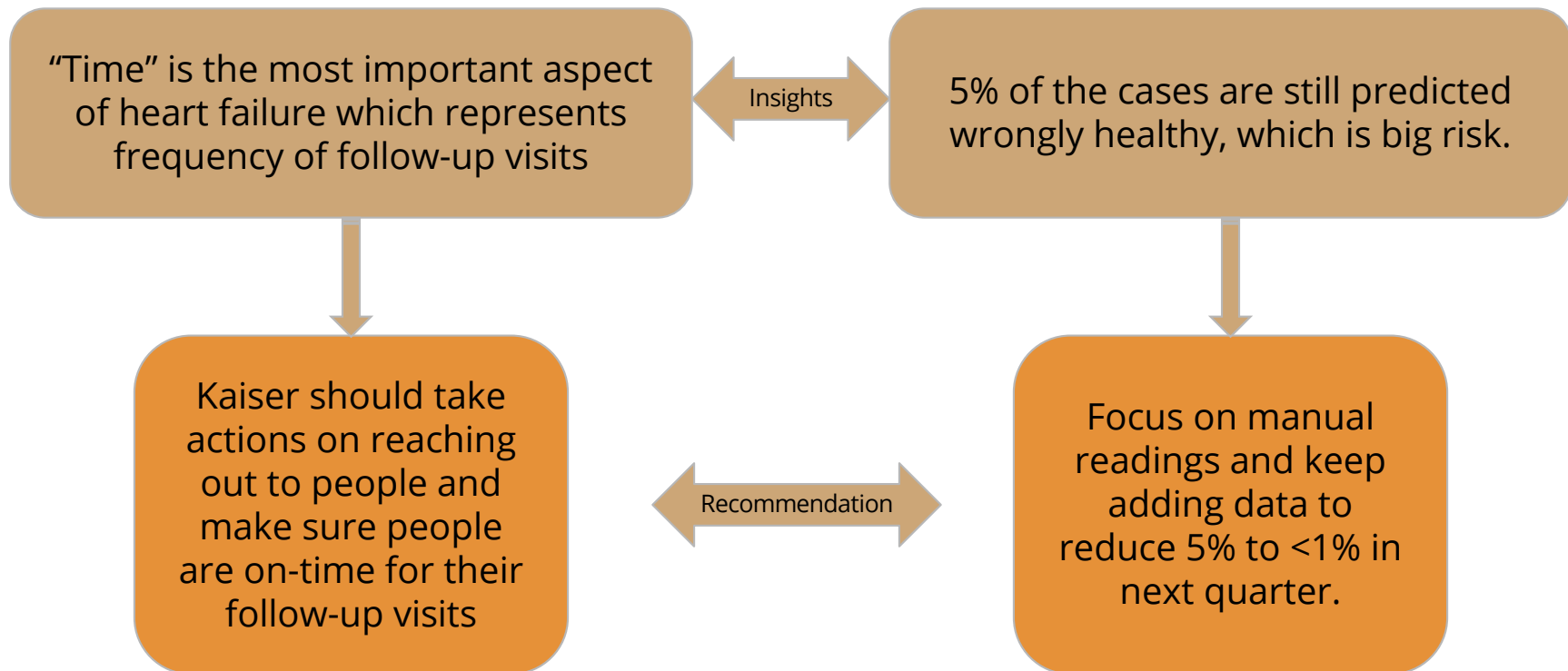
- **43 True Negatives:** Correctly identified as no heart issue.
- **14 True Positives:** Correctly identified as at-risk.
- **3 False Negatives:** Missed at-risk cases.
- **0 False Positives:** No healthy cases mislabeled as at-risk.



# Recommendation & Future Work



# Now what?



# What are our next steps?

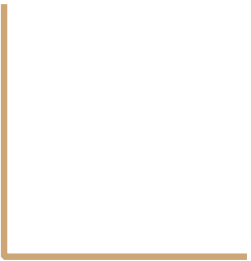
- **Adding more data:** 299 patients data was used in this analysis, as we add the data, better the prediction gets.
  - **Using better Models:**
    - 1st Option: Team can tune the model to give better recommendation
    - 2nd Option: Team can experiment with different ML algorithms to check if prediction gets better.
-

Thank you!





# Appendix



# Data Description

The dataset contains the following columns:

1. Age : Age of the patient
  2. Anaemia : Whether the patient has anaemia (0 for No, 1 for Yes)
  3. creatinine\_phosphokinase: Level of the enzyme in the blood
  4. diabetes: Whether the patient has diabetes (0 for No, 1 for Yes)
  5. ejection\_fraction: Percentage of blood leaving the heart during each contraction
  6. high\_blood\_pressure: Whether the patient has high blood pressure (0 for No, 1 for Yes)
  7. platelets: Quantity of platelets in the blood
  8. serum\_creatinine: Level of serum creatinine in the blood
  9. serum\_sodium: Level of serum sodium in the blood
  10. sex: Gender of the patient (presumably 0 for Female, 1 for Male)
  11. smoking: Whether the patient smokes (0 for No, 1 for Yes)
  12. time: Time (not specified if it's in days, weeks, or months)
  13. criticality\_rank: Ranking of the patient's criticality (higher rank presumably indicates higher risk) **(created for this project)**
  14. DEATH\_EVENT: Whether the patient had a heart failure event (0 for No, 1 for Yes)
-

# Detailed Overview of Our Random Forest Classifier for Heart Failure Prediction

- Our model is based on the Random Forest Classifier, a robust machine learning algorithm.
  - The dataset comprises various clinical and demographic features: age, presence of anemia, creatinine phosphokinase levels, diabetes status, ejection fraction, high blood pressure, platelet counts, serum creatinine and sodium levels, sex, smoking status, follow-up period ('time'), and a criticality rank.
  - The model works by constructing numerous decision trees during training. Each tree independently assesses the data, making a prediction. The model's final output is the majority vote of these trees.
  - Key parameters of the Random Forest, like the number of trees, depth of trees, and criteria for splitting, were optimized based on cross-validation to ensure robust performance.
  - The model's high accuracy (95%) indicates its effectiveness in distinguishing between patients at risk of heart failure and those not at risk.
  - This accuracy was computed using standard metrics like the accuracy score, comparing the model's predictions against actual outcomes in a test dataset.
  - The model's reliability and accuracy make it a valuable tool for early identification of patients at risk of heart failure, allowing for timely intervention and management.
-