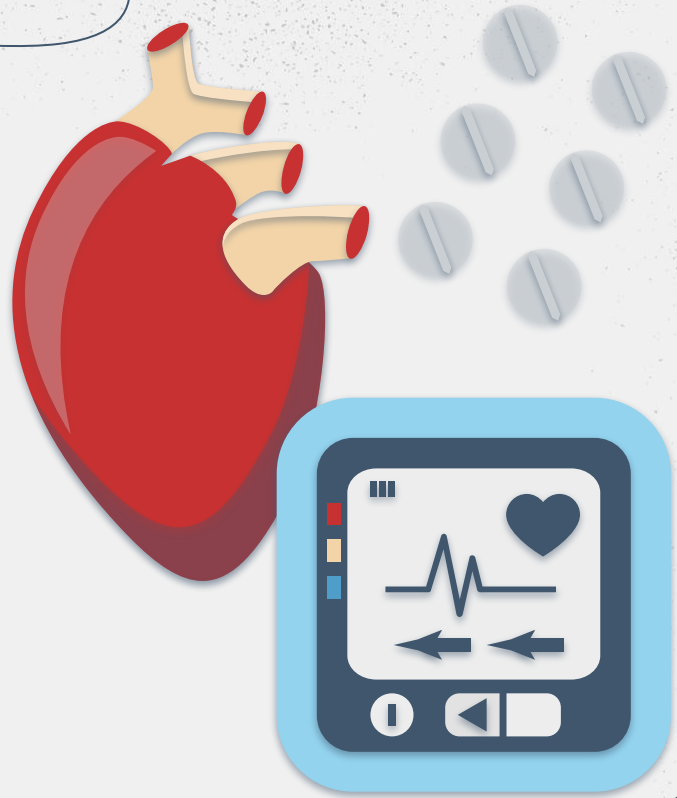
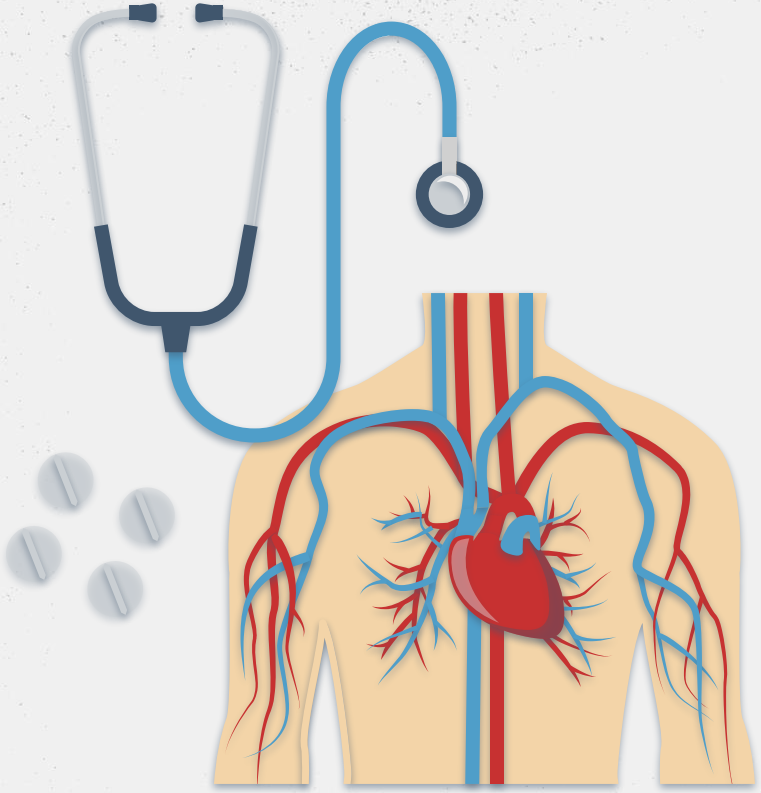
Cardiac Arrest Prediction Using Machine Learning



STAT 5000: Statistical Methods and Applications I

Alex Yarosh

TABLE OF CONTENTS



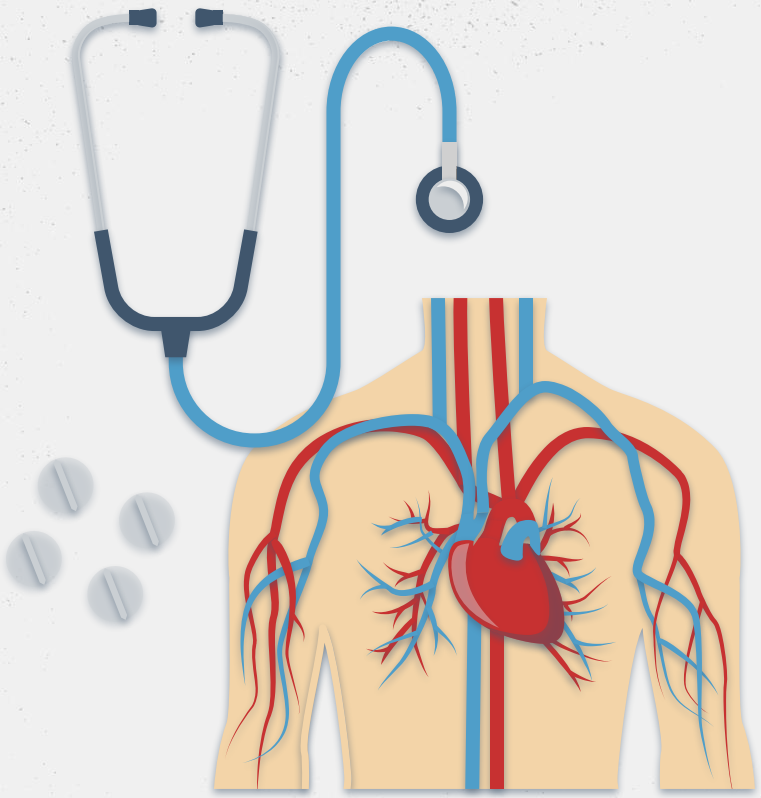
01 INTRODUCTION

Problem Statement

- **Challenge –**
  - Did you know that cardiac arrest claims the lives of over 350,000 Americans annually? It can strike anyone, anytime, often without warning
  - Early detection and intervention are crucial in improving survival rates
  - Existing methods rely on subjective assessments and lack accuracy
- **Consequences –**
  - Delayed diagnosis can lead to irreversible brain damage or death
  - The current approach to cardiac arrest is reactive, not proactive
- **Goal –**
  - To implement a Machine Learning model that can accurately predict cardiac arrest risk based on readily available patient data

About the Dataset

- **Data Source –**
  - It is curated from the **National Health and Nutritional Examination Survey** (NHANES)
  - It has 51 features and 37,079 records
- **Data Contains –**
  - Demographic, socio-economic, health history, blood test, lifestyle, and disease information of individuals, including presence of coronary heart disease.
- **Data Challenges –**
  - There was high class imbalance in the dataset

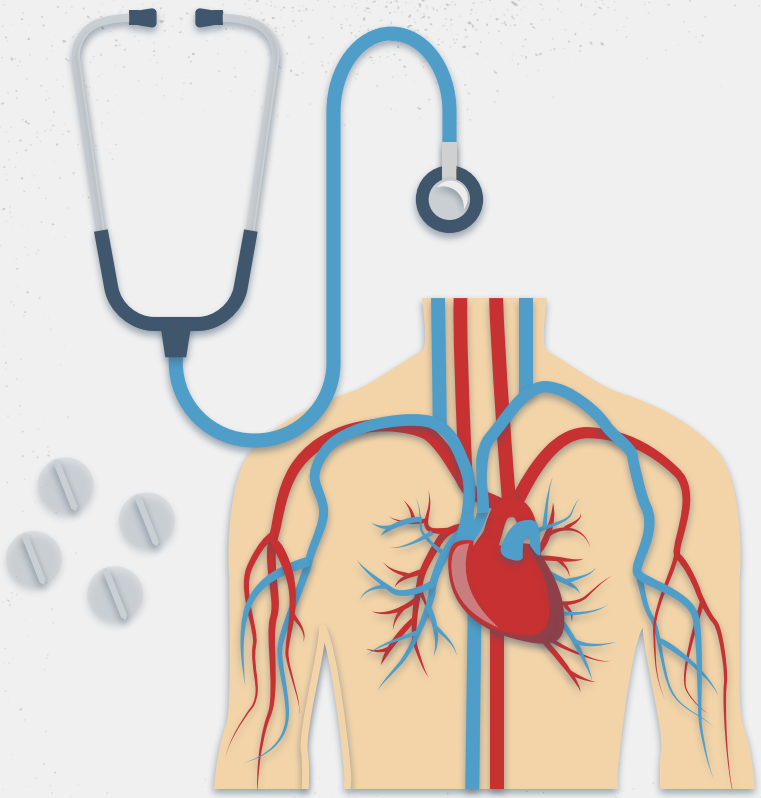


02

DATA PREPARATION

Dataset Preparation

- **Data Cleaning -**
  - Found that there are no missing values
  - There are some outliers
- **Data Balancing -**
  - Resampled the imbalanced target variable by under-sampling majority class (10:1) while artificially increasing the minority class by over-sampling using SMOTE via pipeline
- **Why SMOTE** (Synthetic Minority Oversampling Technique) **-**
  - SMOTE balances imbalanced data by generating synthetic samples for the minority class without discarding real data, potentially preserving valuable information while improving model performance
  - Compared to simpler oversampling methods like duplication, SMOTE reduces overfitting risks and potentially yielding more generalizable models



03

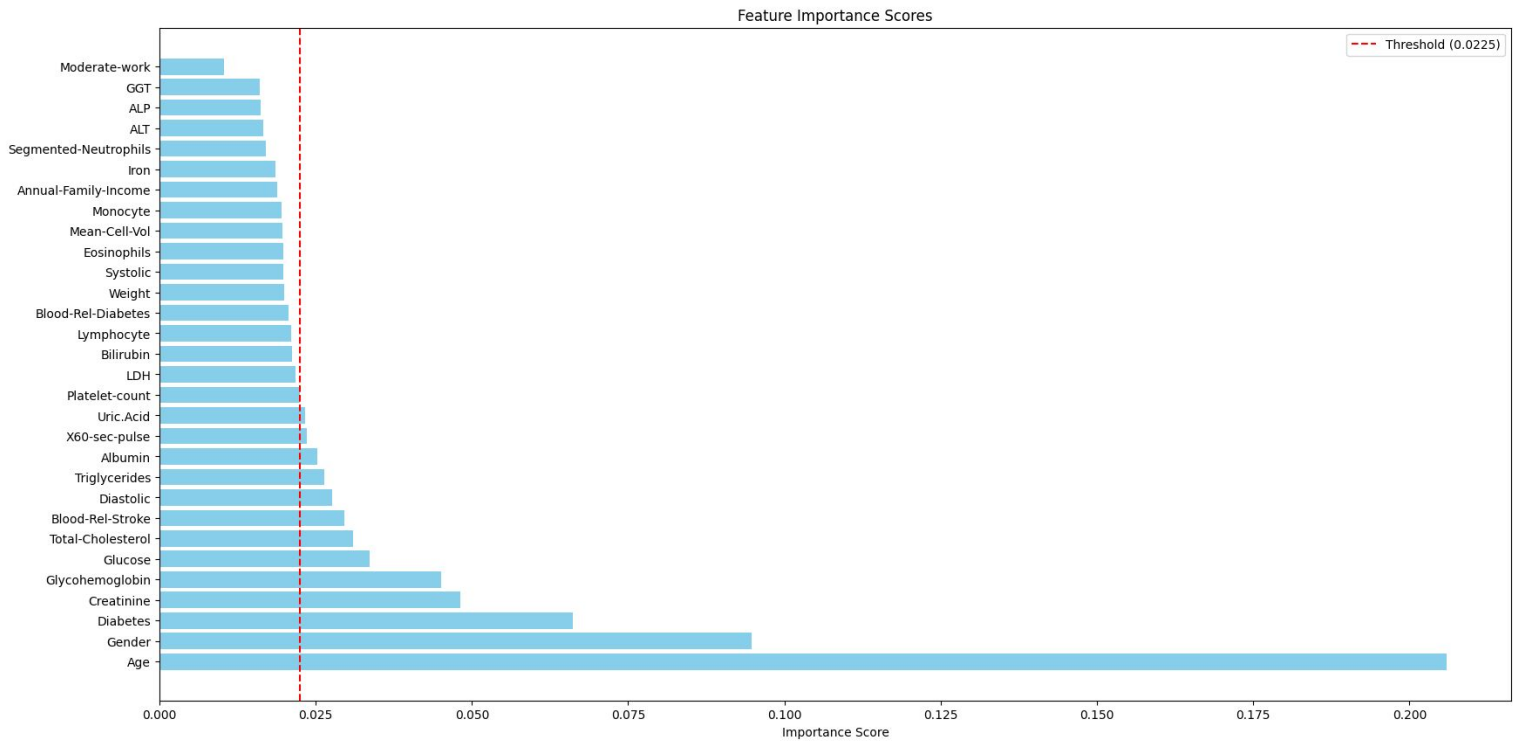
FEATURE SELECTION

Feature Selection

- **Importance -**
  - Not all features are equally informative for prediction.
  - We need to identify the most relevant ones to avoid overfitting and improve model interpretability.

- **Feature Importance Methods -**
  - Different types are Lasso Regression, Chi-squared test, ANOVA, Recursive Feature Elimination, Select K Best, Random Forest.
  - We used **Select K Best + Random Forest.** SelectKBest with Chi-Squared test function prunes irrelevant features, making Random Forest search easier.
  - Random Forest refines top contenders from SelectKBest, revealing key predictors.

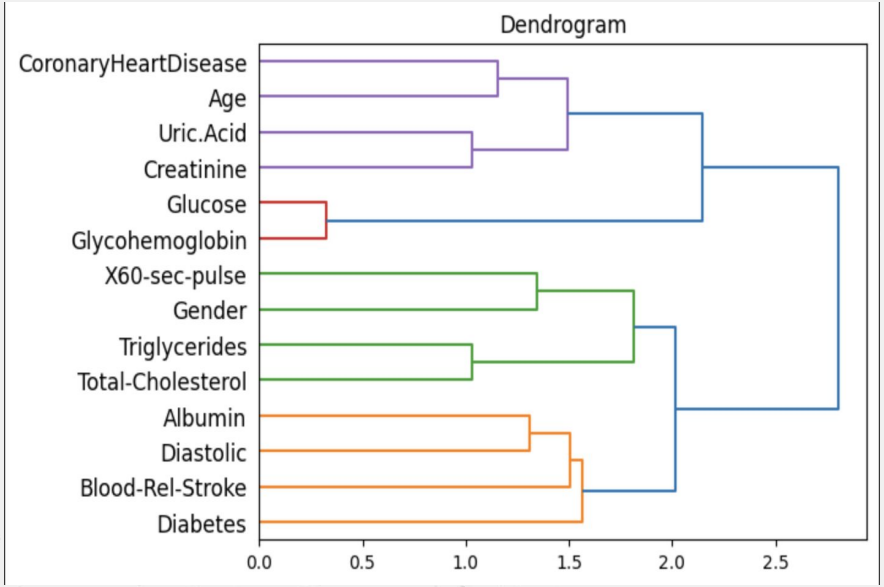
Feature Selection



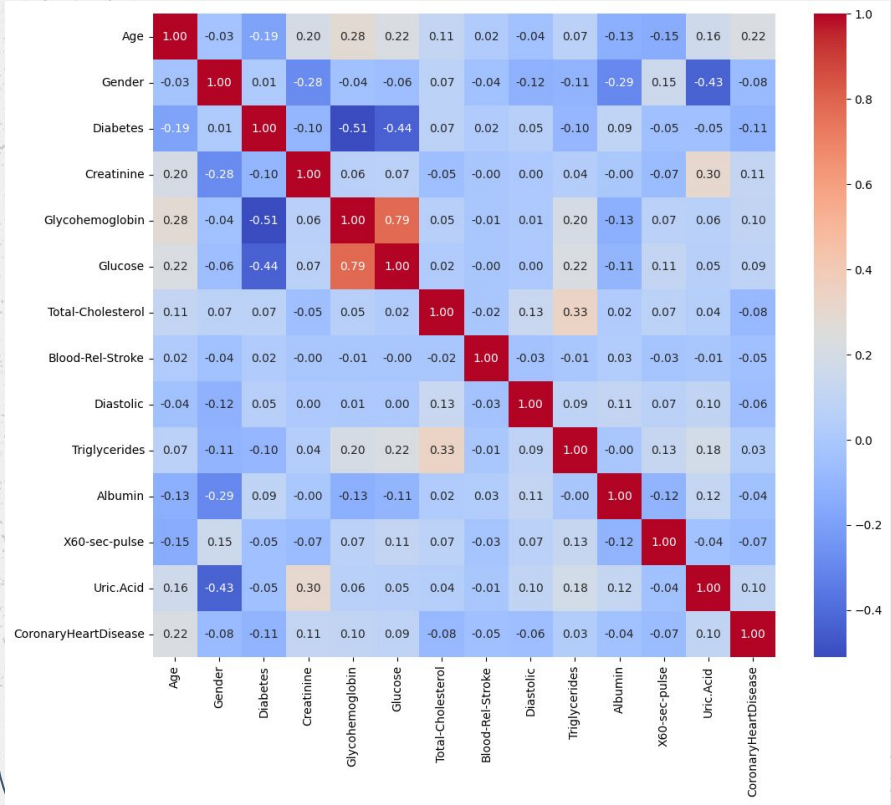
DENDOGRAM OF THE EXTRACTED FEATURES

Top Features Extracted -

'Age', 'Gender', 'Diabetes', 'Creatinine', 'Glycohemoglobin', 'Glucose', 'Total-Cholesterol', 'Blood-Rel-Stroke', 'Diastolic', 'Triglycerides', 'Albumin', 'X60-sec-pulse', 'Uric.Acid'

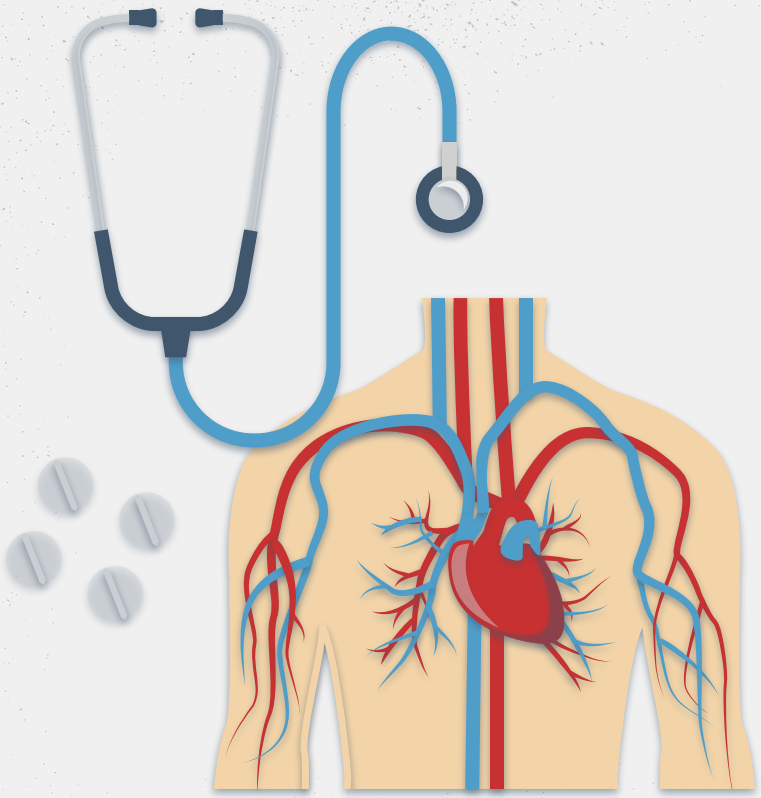


HEATMAP OF THE SELECTED FEATURES



- The Dendogram pointed us towards potentially interesting feature connections & this Heat Map confirms and visualizes these correlations, with warmer colors representing positive associations and cooler colors indicating negative ones

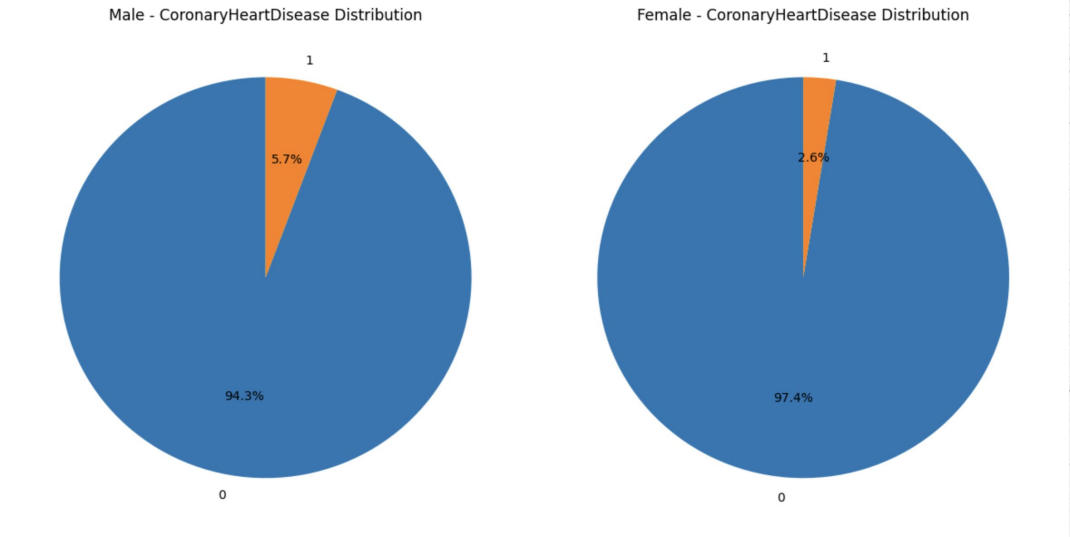
- Set 1 - 'Diabetes', 'Blood-Rel-Stroke', 'Diastolic', 'Albumin'
- Set 2 - 'Glycohemoglobin', 'Glucose'
- Set 3 - 'X60-sec-pulse', 'Gender'
- Set 4 - 'Uric.Acid', 'Creatinine', 'Age', 'Triglycerides', 'Total-Cholesterol'



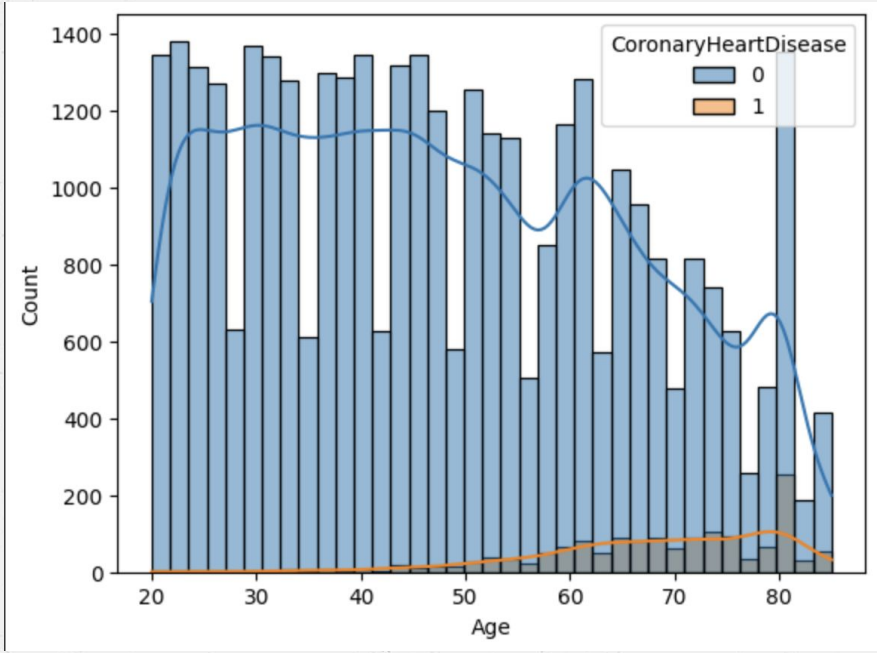
04 EXPLORATORY DATA ANALYSIS

GENDER DISTRIBUTION USING PIE CHART

- Males exhibit a statistically significant trend towards experiencing CHD more frequently than females
- Differences in risk factor prevalence between genders can be noteworthy, with males being more likely to have high cholesterol and smoking habits associated with CHD



HISTOGRAM OF AGE AND COUNT

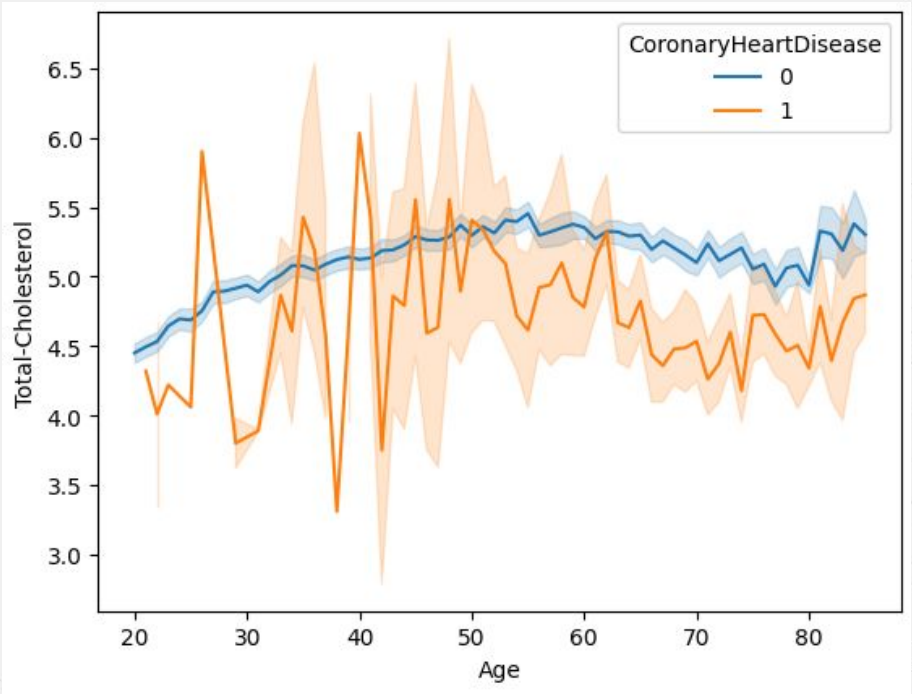


- CHD climbs with age! The graph shows a rising trend, with risk escalating after 50-60 age
- Remember, individuals vary, and other factors besides age play a role

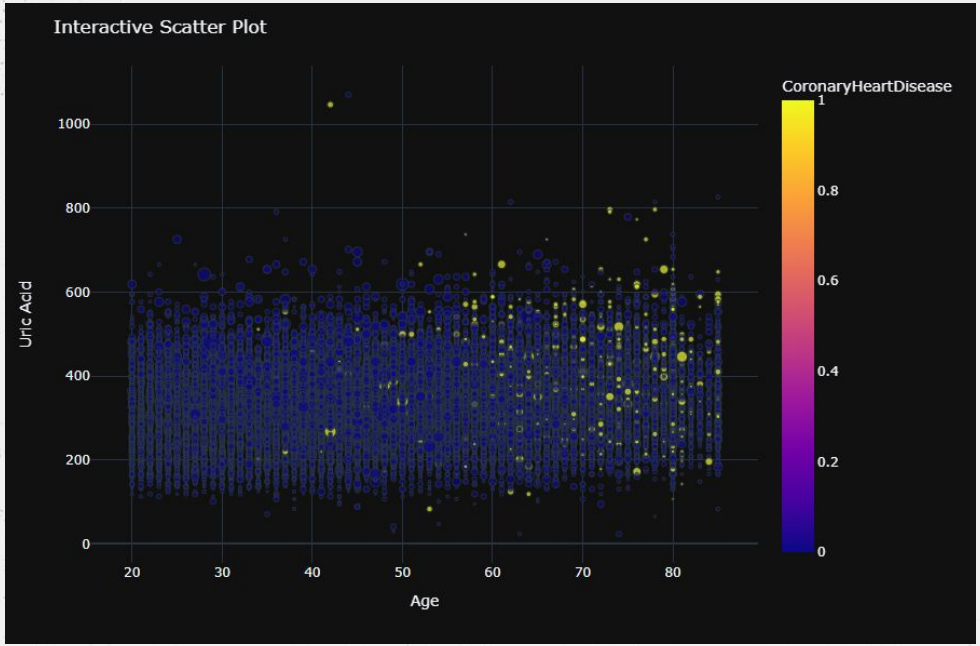
LINE PLOT FOR CHOLESTEROL DISTRIBUTION

- Individuals with higher or lower cholesterol levels than the recommended range tend to have a greater potential for developing CHD

- People with consistently normal cholesterol levels within the healthy range generally face a significantly lower risk of CHD

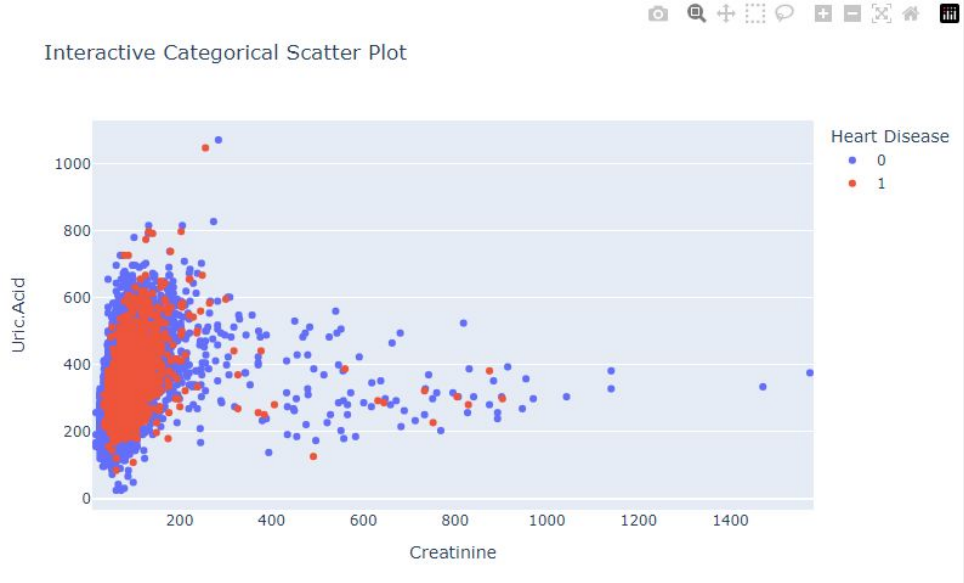


URIC ACID LEVELS USING BUBBLE PLOT

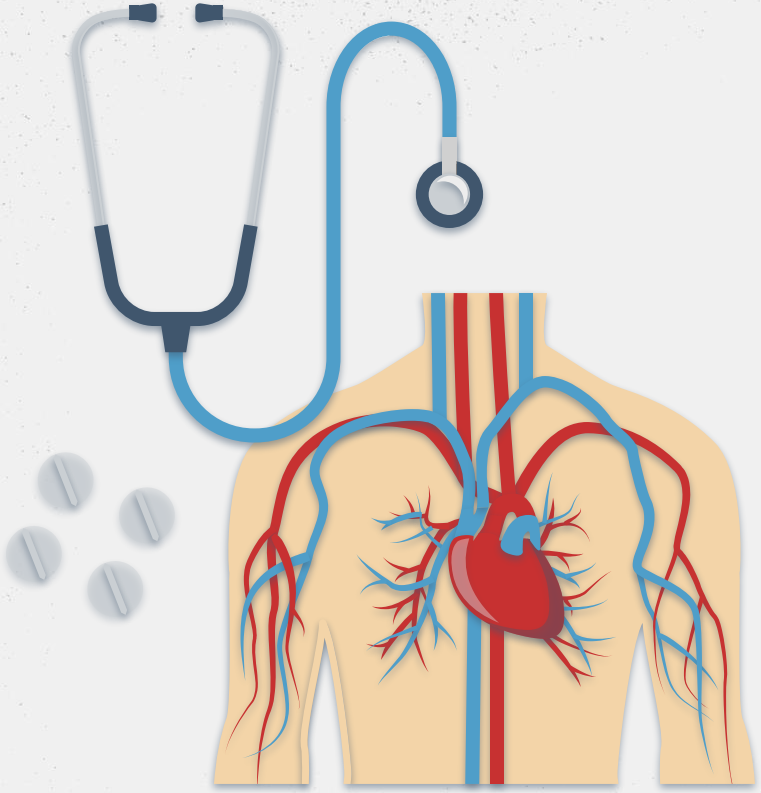


- Imagine a ticking time bomb: with age, levels of triglycerides and uric acid tend to climb (like the fuse burning)
- This can eventually trigger the explosion of coronary heart disease, potentially leading to cardiac arrest

RELATIONSHIP BETWEEN CREATININE & URIC ACID USING SCATTER PLOT



- The relationship between creatinine and uric acid seems complex and non-linear
- While there's a general trend of uric acid increasing with creatinine, individual data points deviate from this trend, highlighting the influence of other factors



05

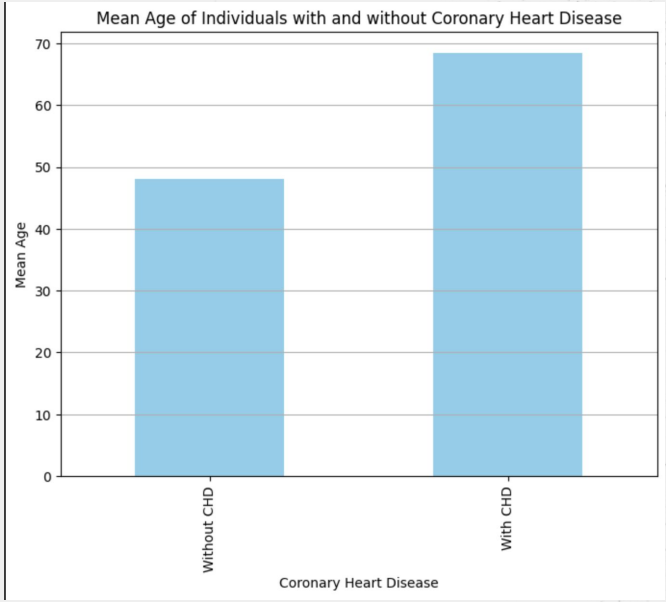
HYPOTHESIS TESTING

HYPOTHESIS TESTING 1

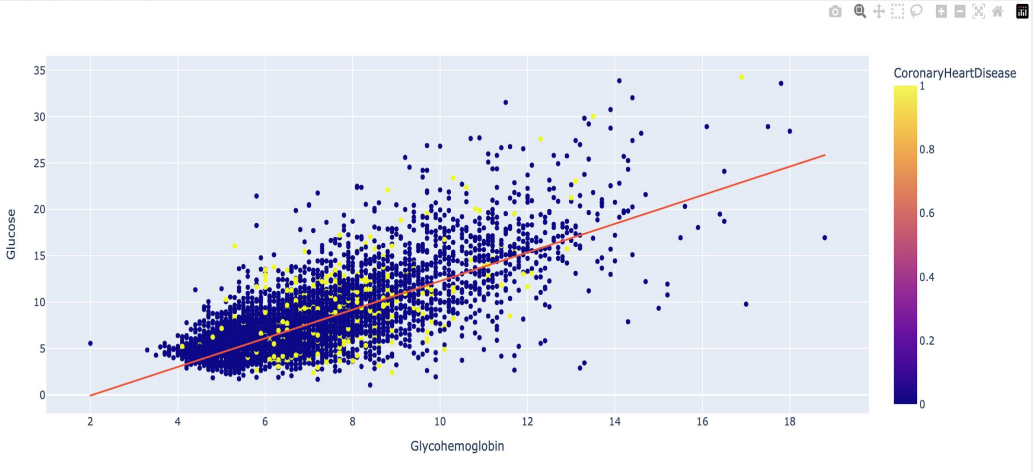
Null Hypothesis (H0): The mean age of individuals with CoronaryHeartDisease is equal to the mean age of individuals without CoronaryHeartDisease

Alternative Hypothesis (H1): The mean age of individuals with CoronaryHeartDisease is different from the mean age of individuals without CoronaryHeartDisease

Result: Reject the null hypothesis. The mean age is significantly different between individuals with and without CoronaryHeartDisease



HYPOTHESIS TESTING 2



Null Hypothesis (H0): The mean Glycohemoglobin level is equal in individuals with and without CoronaryHeartDisease

Alternative Hypothesis (H1): The mean Glycohemoglobin level is different in individuals with and without CoronaryHeartDisease

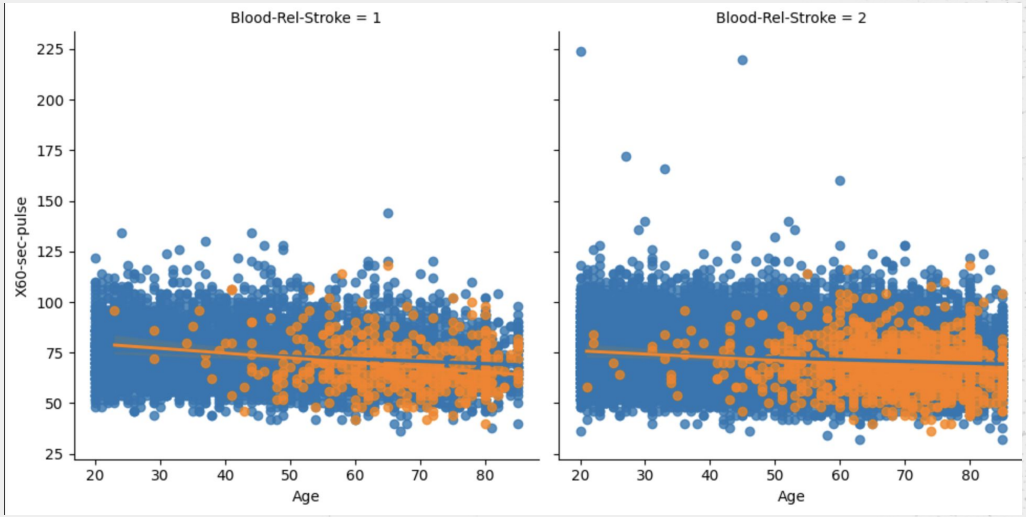
Result: Reject the null hypothesis. The mean Glycohemoglobin level is significantly different between individuals with and without CoronaryHeartDisease

HYPOTHESIS TESTING 3

Null Hypothesis (H0): Having a history of Coronary Heart disease does not have any affect on the pulse rate of people

Alternate Hypothesis (H1): Having a history of Coronary Heart disease has an affect on the pulse rate of people

Result: The Null Hypothesis is rejected, Having a history of Coronary Heart disease has an affect on the pulse rate of people.

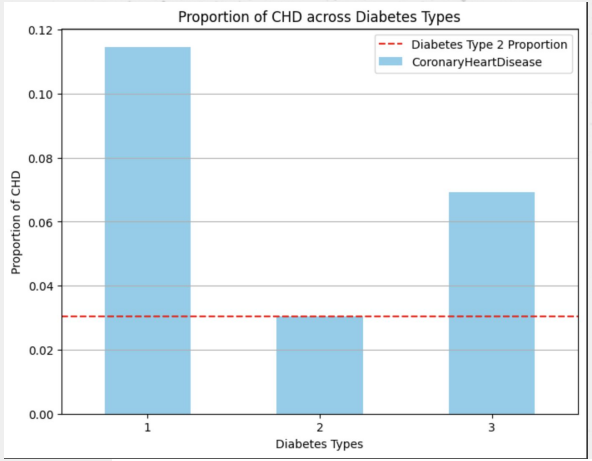


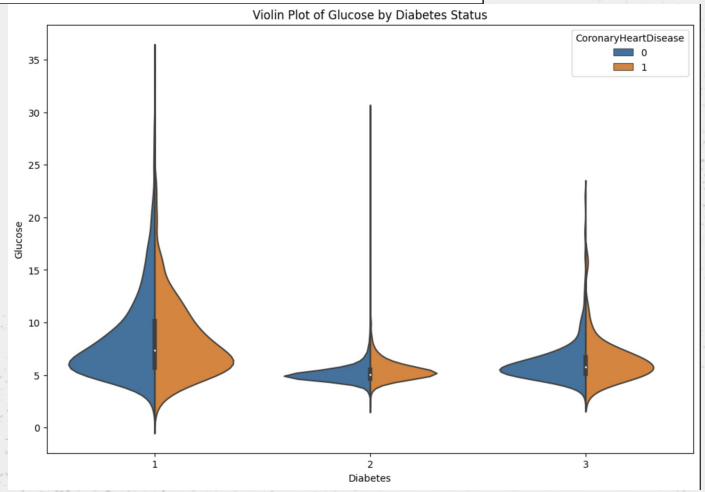
HYPOTHESIS TESTING 4

Null Hypothesis (H0): The proportion of people suffering with CoronaryHeartDisease is same across all the 3 diabetes types

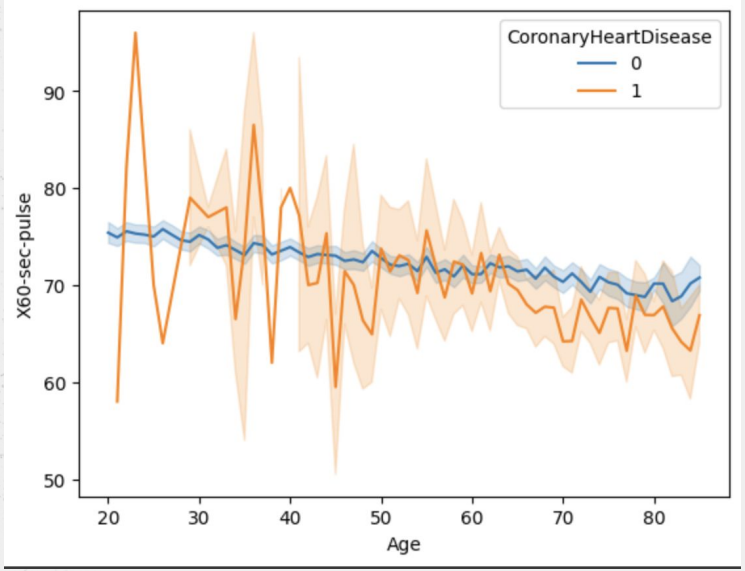
Alternate Hypothesis (H1): The proportion of people suffering with CoronaryHeartDisease is higher with Diabetes type 2 patients than the rest

Result: Reject the null hypothesis. There is evidence to suggest that the proportion of people suffering from CoronaryHeartDisease is different across diabetes types.





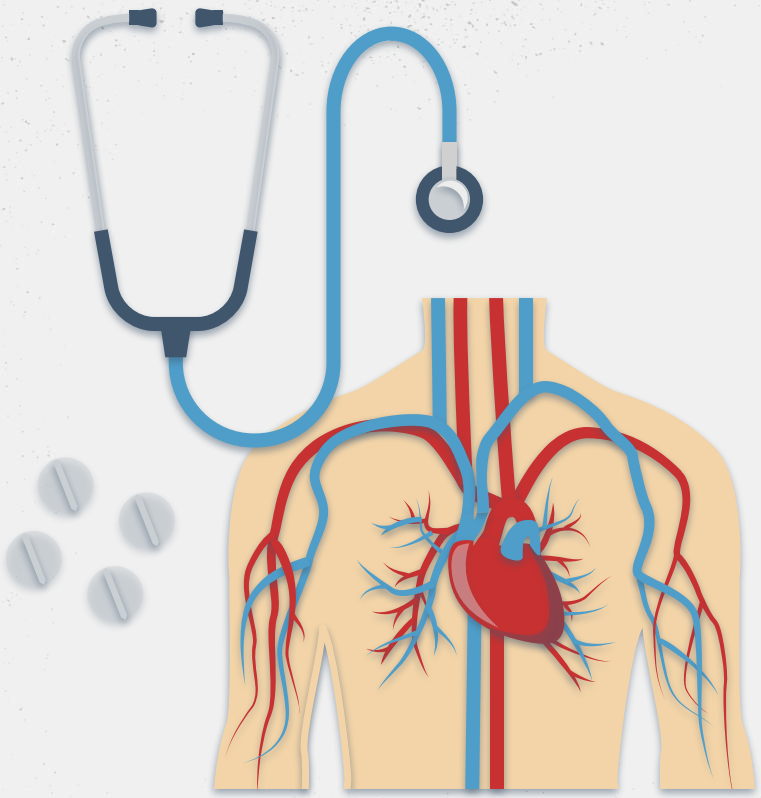
HYPOTHESIS TESTING 5



Null Hypothesis (H0): The pulse rate varies the same for people with and without Coronary Heart Disease

Alternative Hypothesis (H1): The pulse rate varies the rigorously for people with Coronary Heart Disease than people without any heart disease

Result: Reject the null hypothesis. There is evidence to suggest that the pulse rate varies rigorously for people with CoronaryHeartDisease than people without any heart disease



06

MODEL TRAINING & EVALUATION

Model Used for Training

❖ Support Vector Machine (SVM) :

Powerful classifiers finding maximum margin hyperplanes between classes

❖ Random Forest :

Ensemble learner building diverse decision trees for robust predictions

❖ Logistic Regression :

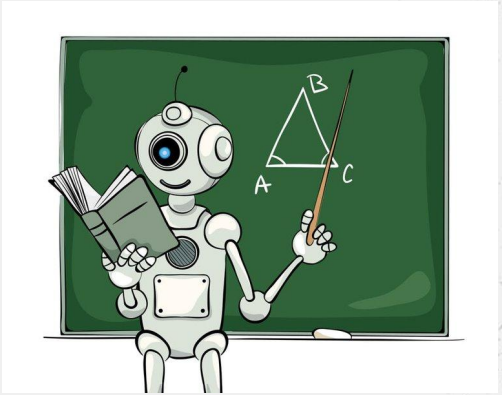
Probabilistic model for binary classification based on linear relationships

❖ Gradient Boosting :

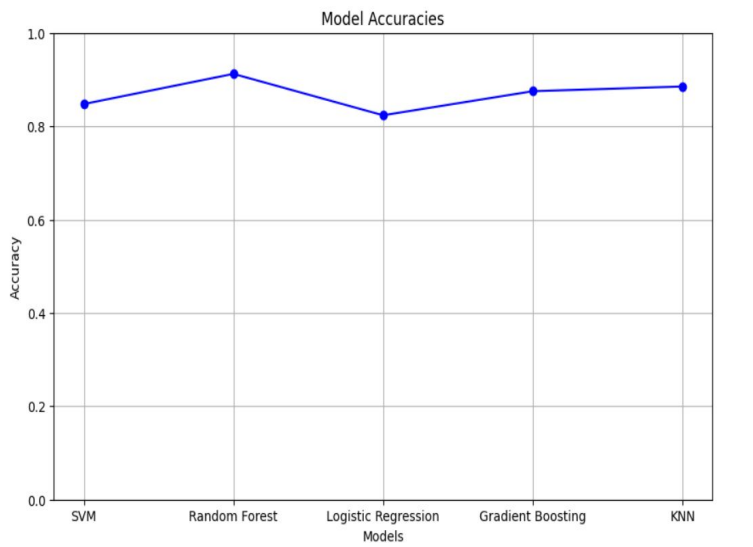
Sequential ensemble learner iteratively improving predictions on residuals

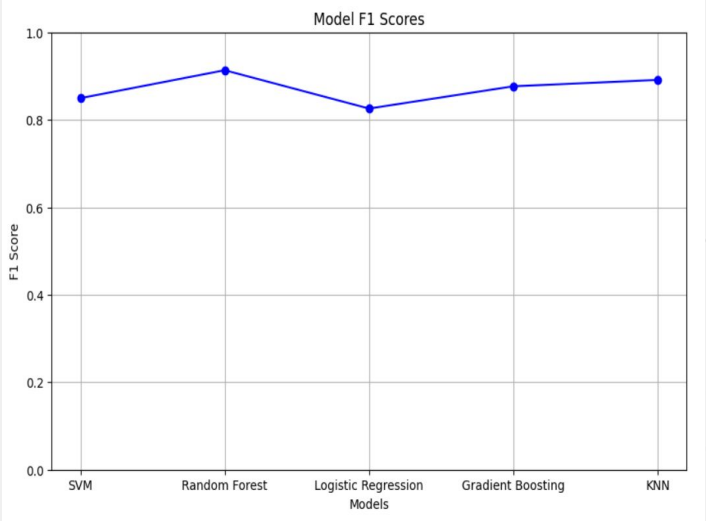
❖ K-Nearest Neighbors Algorithm (KNN) :

Classifies points based on distance to nearest neighbors in the training data



Models Evaluation

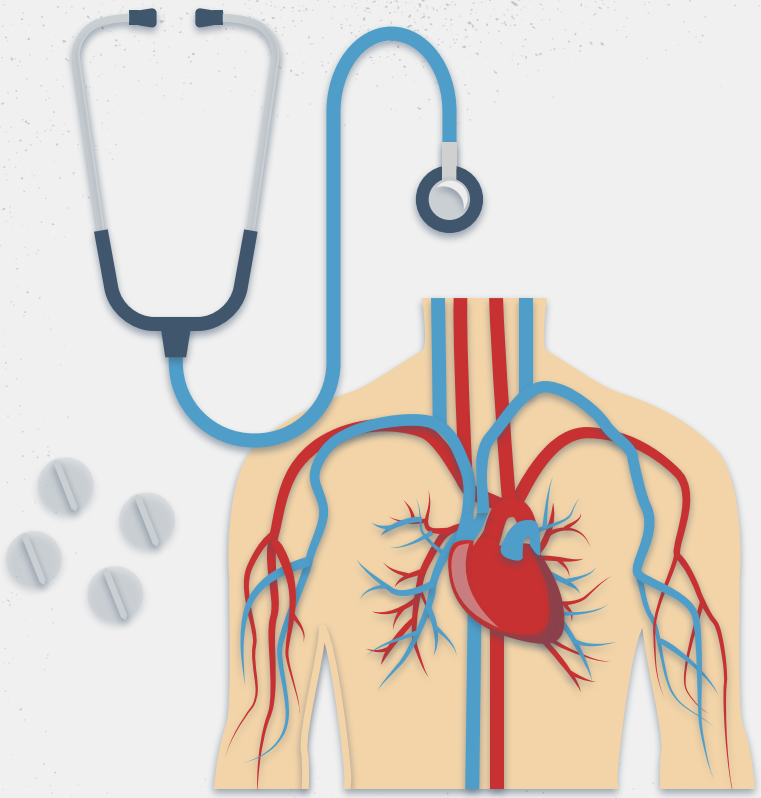




Key Findings

Models	Accuracy	F1-score	AUC score
Support Vector Machine (SVM)	84.74%	0.8495	0.8480
Random Forest	91.21%	0.9133	0.9127
Logistic Regression	82.34%	0.8257	0.8239
Gradient Boosting	87.5%	0.8766	0.8755
K-Nearest Neighbors Algorithm (KNN)	88.51%	0.8912	0.8864

Random Forest reigns supreme! Its 91.2% accuracy, coupled with strong F1 and AUC scores
Even Gradient Boosting, at 87.5% accuracy, falls short in precision and balance



07

CONCLUSION

Future Work & Recommendations

- **Improvements -**
  - We plan to incorporate additional data sources like ECG recordings or Medical images to further refine the model's accuracy
- **Real-World Applications -**
  - This model could be integrated into Electronic Health Records systems to provide real-time risk prediction at the point of care
- **Call to Action -**
  - By leveraging the power of Machine Learning, we can revolutionize cardiac arrest prediction and improve patient outcomes
  - Let's continue to invest in research and development in this

THANK YOU !

Presented By

Srimedha Bhavani Chandoo
Jagrati Chauhan
Sai Pratheek KVDSNK
Navya Prasad Malur Narasimha Prasad
Sai Krishna Sriram