# Predicting Coronary Heart Disease Risk Using the Framingham Heart Study Dataset

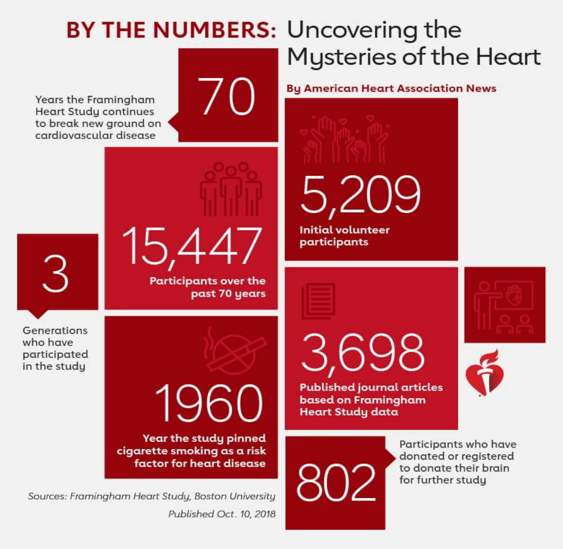
Vaishali Sharma
Data Science Career Track

SCHOOL OF DATA
by Springboard

# Major global health concerns:
## High cholesterol and CHD

**Objective**: Develop a predictive model to estimate the likelihood of coronary heart disease (CHD) using health and lifestyle data.

**Dataset**: Framingham Heart Study, including variables like cholesterol, blood pressure, age, smoking, and family history.



**BY THE NUMBERS:** Uncovering the Mysteries of the Heart

By American Heart Association News

70 Years the Framingham Heart Study continues to break new ground on cardiovascular disease

5,209 Initial volunteer participants

15,447 Participants over the past 70 years

3 Generations who have participated in the study

3,698 Published journal articles based on Framingham Heart Study data

1960 Year the study pinned cigarette smoking as a risk factor for heart disease

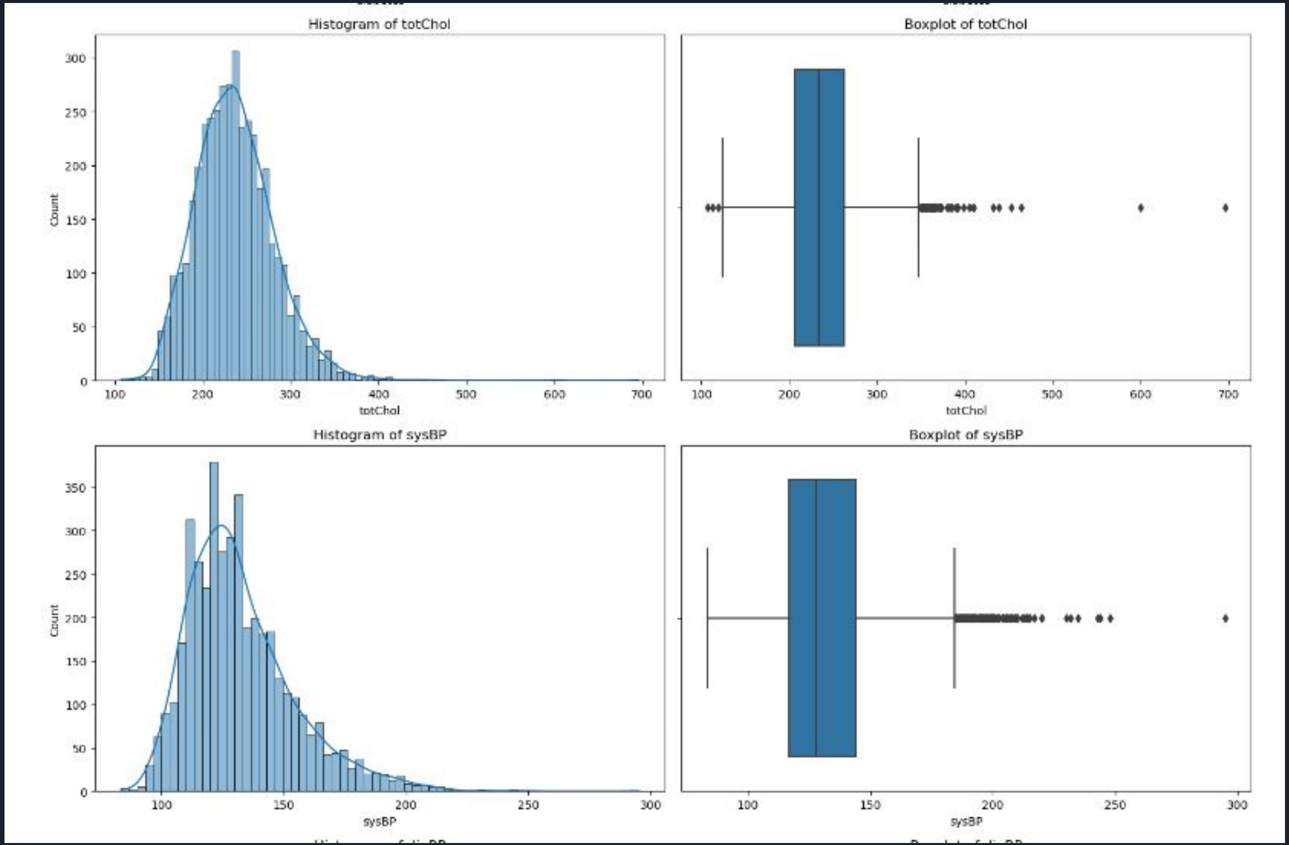
802 Participants who have donated or registered to donate their brain for further study

*Sources: Framingham Heart Study, Boston University*

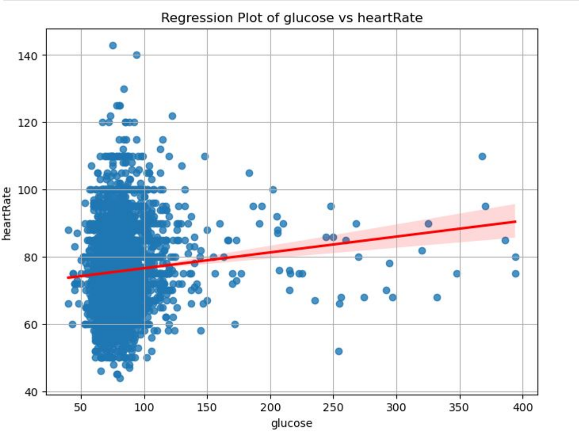
*Published Oct. 10, 2018*

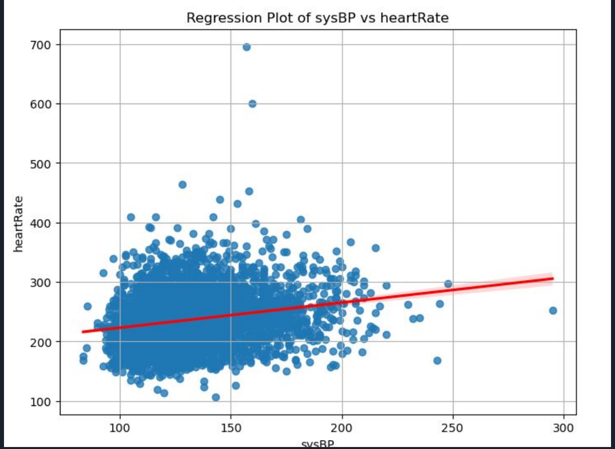
https://www.heart.org/en/news/2018/10/10/framingham-the-study-and-the-town-that-changed-the-health-of-a-generation

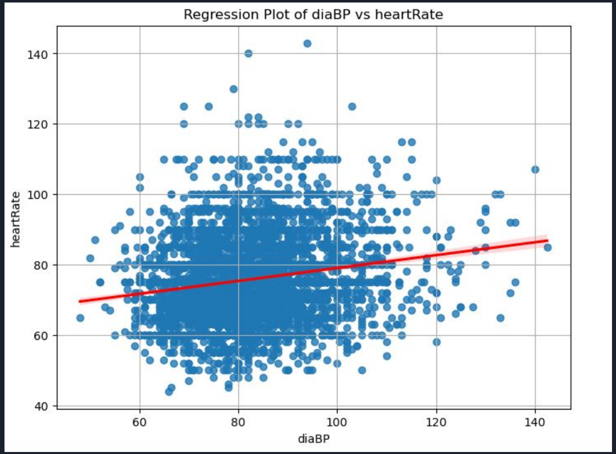
# Understanding dataset



# Trends and relationships







# Processing Data

Handled missing values and normalized numerical features.

Encoded categorical variables.

Split the data into training and testing sets.

# Machine Learning : Modeling

**Resampling the unbalanced dataset**

**Machine Learning**:

- **Logistic Regression** (Baseline model)
- **Random Forest** (Ensemble method)
- **XGBoost** (Gradient boosting)

# Model Comparison

**Logistic Regression**:

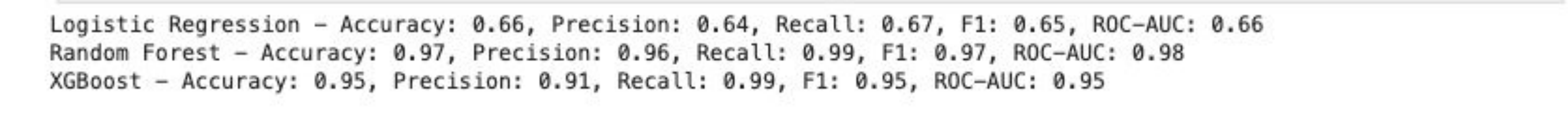
- Moderate performance but low accuracy and ROC-AUC.

**Random Forest**:

- Best performance, with high accuracy, precision, recall, and F1 score.

**XGBoost**:

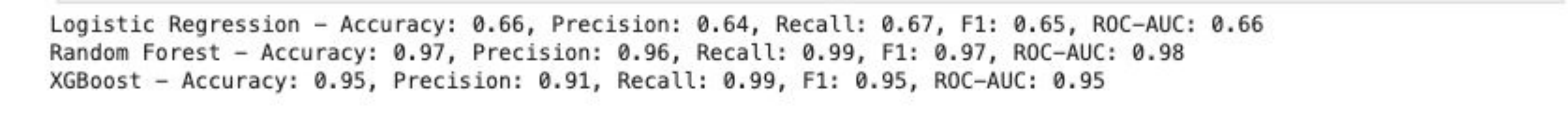
- Strong performance, close to Random Forest but slightly lower precision.



```
Logistic Regression - Accuracy: 0.66, Precision: 0.64, Recall: 0.67, F1: 0.65, ROC-AUC: 0.66
Random Forest - Accuracy: 0.97, Precision: 0.96, Recall: 0.99, F1: 0.97, ROC-AUC: 0.98
XGBoost - Accuracy: 0.95, Precision: 0.91, Recall: 0.99, F1: 0.95, ROC-AUC: 0.95
```

# Results & Inference

- **Best Model**: Random Forest – highest overall performance across accuracy, recall, and precision.
- **XGBoost**: Close second, strong alternative for faster computations.
- **Logistic Regression**: Falls short in comparison to ensemble methods.



```
Logistic Regression - Accuracy: 0.66, Precision: 0.64, Recall: 0.67, F1: 0.65, ROC-AUC: 0.66
Random Forest - Accuracy: 0.97, Precision: 0.96, Recall: 0.99, F1: 0.97, ROC-AUC: 0.98
XGBoost - Accuracy: 0.95, Precision: 0.91, Recall: 0.99, F1: 0.95, ROC-AUC: 0.95
```

Random Forest : most reliable model for predicting CHD in this dataset. (amongst the models utilized)

Provides actionable insights for healthcare professionals to assess and manage CHD risk.

# Thank you!