Project Background

•••

The Heart Disease Dataset

Agenda

1) Problem Statement
2) The Importance of this Research
3) The Dataset
4) Previous Models and Research
5) Research Gaps
6) Our Contribution and Next Steps

Problem Statement

Goal: Develop machine learning models to predict future heart disease occurrences

- Analyze significance of predictive factors
- Utilize machine learning to create a predictive model
- Implement model in a clinical environment

Importance

Heart Disease:

- Leading cause of death in both men and women of most racial and ethnic groups for 100 years (Over 17.9 Million deaths annually worldwide)
- Cost people anywhere from billions to trillions of dollars (US spends $219 billion per year (CDC))
- In the United States a heart attack happens every 40 seconds, 1 in 5 of those are silent
- Early Detection reduces mortality by 30-40%

Sources:

- https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html
- https://newsroom.heart.org/news/more-than-half-of-u-s-adults-dont-know-heart-disease-is-leading-cause-of-death-despite-100-year-reign

The Dataset

Name: Heart Disease Classification:

- 14 Columns (Test Results)

- 302 Rows (Patients)

Variables

Age: Patients age
Sex: Patients gender
cp: Chest Pain Type

- **0:** Asymptomatic (no chest pain)
- **1:** Typical angina (chest pain related to the heart)
- **2:** Atypical angina (chest pain that isn't clearly heart-related)
- **3:** Non-anginal pain (chest pain not related to the heart at all)

trestbps: Resting blood pressure (in mm Hg on admission to the hospital)
chol: Serum cholesterol in mg/dl
fbs: Fasting Blood Sugar (fasting blood sugar > 120 mg/dl)

- 0 = false
- 1 = true

Variables

restecg: Resting electrocardiographic results

- Value 0: normal
- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

thalach: Total Maximum Heart Rate Achieved

exang: Exercise-Induced Angina

- 0 = no
- 1 = yes

oldpeak: ST depression induced by exercise relative to rest - (depression: falling below the baseline)

- ST segment: The period where the ventricles are fully contracted but before they start to relax and "reset" for the next beat.

Variables

slope: Slope of the peak exercise ST segment
- Value 0: upsloping
- Value 1: flat
- Value 2: downsloping

ca: Number of major vessels (0-3) colored by fluoroscopy

thal: Thallium Stress Test Results (?)
- 0 = error (in the original dataset 0 maps to NaN's)
- 1 = fixed defect
- 2 = normal
- 3 = reversible defect

target: Whether or not the subject has Heart Disease

Previous Models

The National Library of Medicine has conducted a previous study on CVD using machine learning techniques

- Used classifiers such as K-NN, Support Vector Machine, Logistic Regression, Random Forest, and more...
- Study concluded that an optimized XGBoost model had the most desirable outcomes
  - 98.5% accuracy
  - 99.1% recall
  - 98.7% F1 score

Source: https://pmc.ncbi.nlm.nih.gov/articles/PMC10813849/#notes4

Past Work — Other Approaches

Logistic Regression :

- Simple interpretable , ~80-85% accuracy

Decision Trees

- Easy to explain, but overfits

Random Forest

- More Robust, Higher Accuracy, Less Interpretable

Neural Networks

- Powerful, but “Black Box” & resource-intensive

Gaps in Past Research

- Small Datasets = Limited generalization
- Focus on accuracy, not interpretability or client usability
- High-Performing models can be too complex for real world use
  - XGBoost, Neural Networks
- Some models require heavy computation
  - Not practical for clinics
- Need for systematic comparison under same conditions

Our Contribution and Next Steps

We plan to:

- Test and compare multiple models
  - Logistic Regression, Decision Tree, Random Forest, Neural Network
- Evaluate accuracy + interpretability side by side
- Explore which health factors matter most in predictions
- Provide insights that can guide clinicians in real world practice

Summary

Overall:

- Heart Disease is urgent and costly
  - Prevention is key
- Machine Learning shows promise for prediction
- Previous work highlights high accuracy but with gaps
- Our project aims to bridge accuracy and interpretability