Predicting Cardiovascular Disease

Alyssa Adamczak, Leah Bishop, Laxmi Sowjanya Doddi, & Jonathan Kivuva

CIS 635

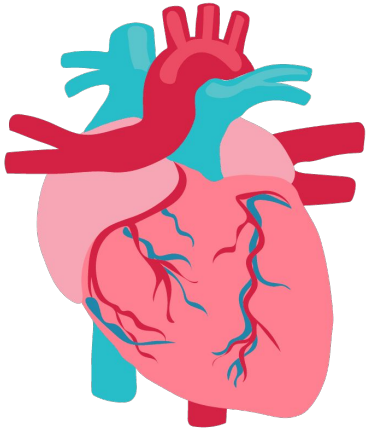
Overview

- Introduction to Cardiovascular Disease
- Our dataset
- Models used
- Results
- Key takeaways

Cardiovascular Disease[1]

- Cardiovascular disease (CVD) is considered the leading cause of death globally

- CVD refers to diseases involving the heart and blood vessels, including:
  - Stroke
  - Heart failure
  - Hypertensive heart disease
  - Rheumatic heart disease
  - Peripheral arterial disease



1. Adhikary D, Barman S, Ranjan R, Stone H. A Systematic Review of Major Cardiovascular Risk Factors: A Growing Global Health Concern. Cureus. 2022 Oct 10;14(10):e30119. Doi: 10.7759/cureus.30119. PMID: 36381818; PMCID: PMC9644238.

Objective

Build a model that predicts the presence or absence of CVD & assess which features are the most significant risk factors of CVD.

Our Dataset

- *Cardiovascular Disease dataset* from kaggle.com
- 70,000 patient records
- 11 features + 1 target
  - Age, height, weight, & gender (objective)
  - Systolic & diastolic blood pressure (exam feature)
  - Cholesterol & glucose levels (exam feature)
  - Smoking, alcohol intake, & physical activity (subjective)
  - **Target variable:** presence or absence of CVD

Issues with Dataset

- ‘Age’ feature originally given in units of days

  (ex. Age of 18,393 days)

- Unrealistic/impossible values for numeric features
- Highlights the importance of domain knowledge

Heights between 149cm and 171cm (approx. 4'11 to 7'0)

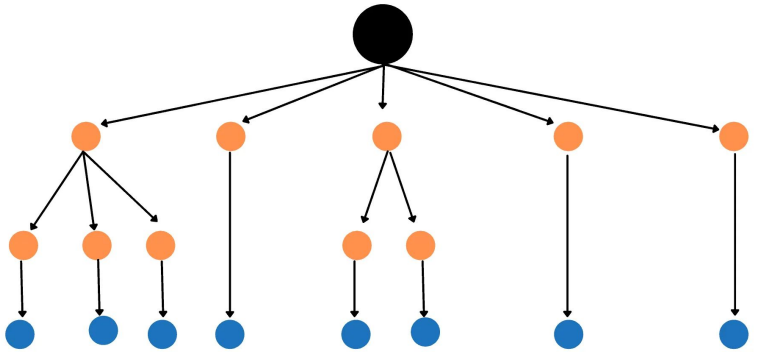
Weights between 38kg and 125kg (approx. 85lbs to 275lbs)

Systolic blood pressures (ap_hi) between 90mmHg and 180mmHg

Diastolic blood pressures (ap_lo) beween 60mmHg and 120mmHg

Models Used

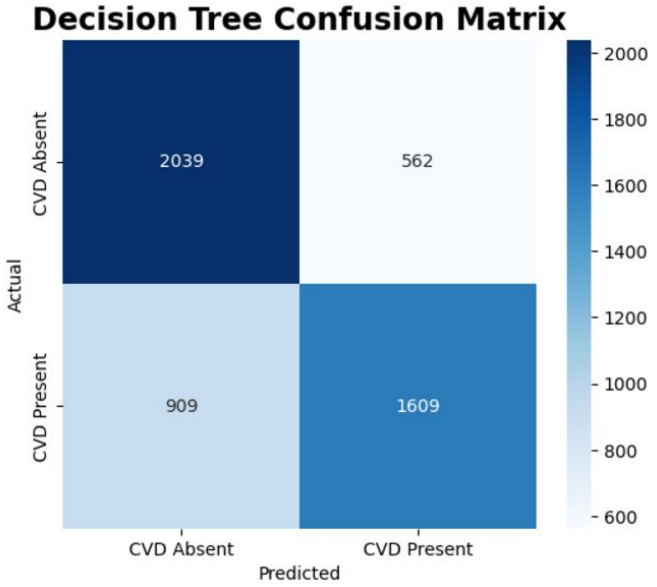
- Decision Tree
- Random Forest
- Logistic Regression



Decision Tree

- A popular supervised machine learning algorithm works by modeling decisions as a tree-like structure

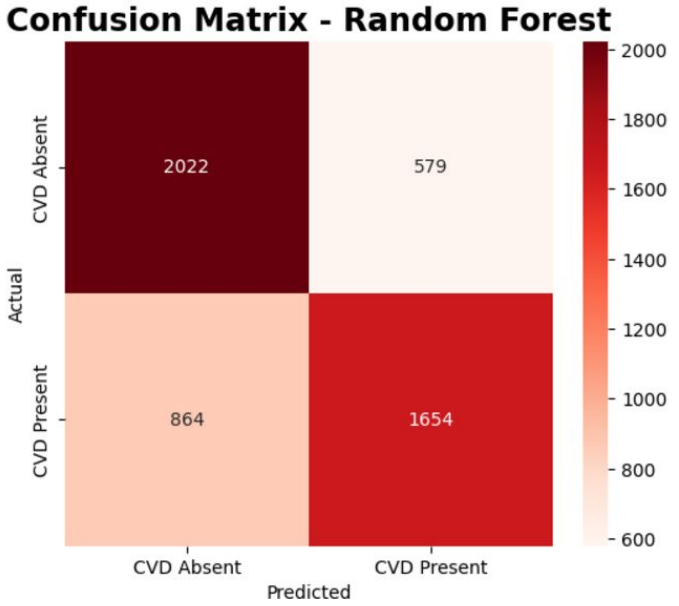
- The performance of the decision tree model was evaluated using several key metrics.
  - Accuracy : 0.713
  - Precision : 0.741
  - Recall(Sensitivity) : 0.639
  - F1 Score : 0.686
  - AUC Score: 0.779



Random Forest

- Ensemble method combining multiple decision trees for robust predictions

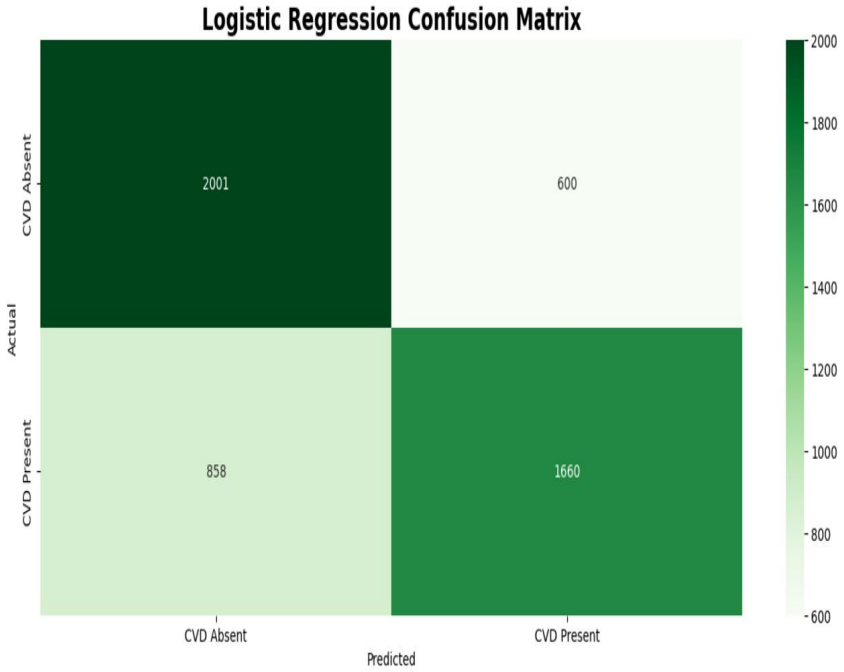
- The random tree's performance was assessed using the following metrics:
  -Accuracy : 0.718
  -Precision : 0.741
  -Recall : 0.657
  -F1 Score : 0.696
  -AUC Score: 0.780



Logistic Regression

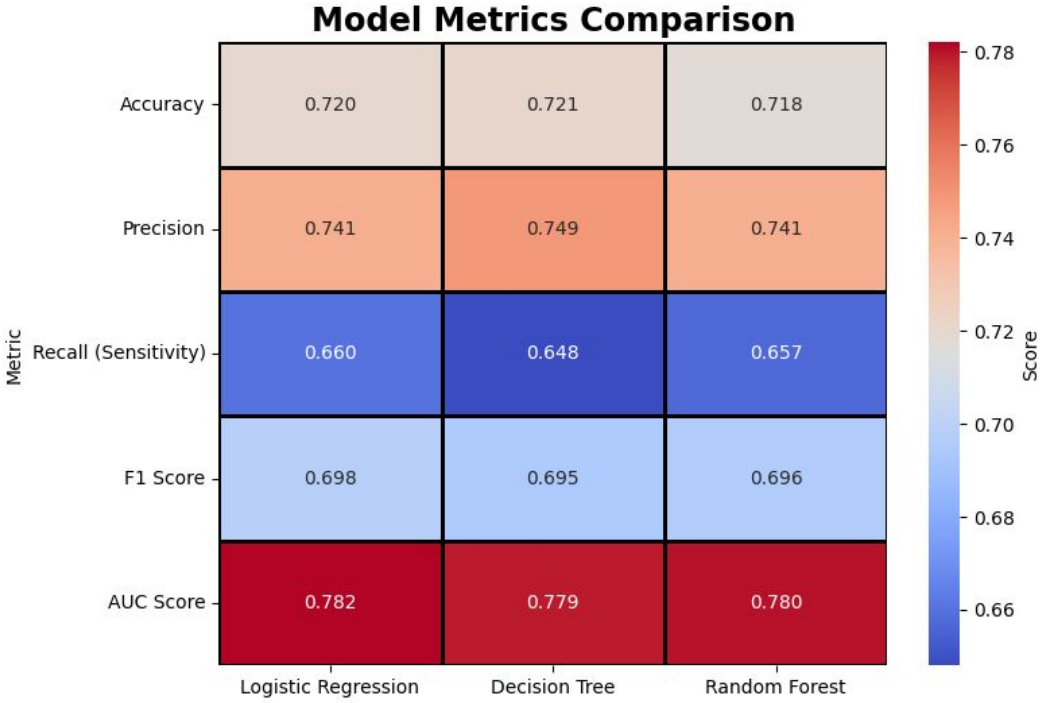
- Linear model for binary classification

- Key performance evaluation metrics :
  - Accuracy: 0.715
  - Precision: 0.735
  - Recall (Sensitivity): 0.659
  - F1 Score: 0.698
  - AUC Score: 0.782



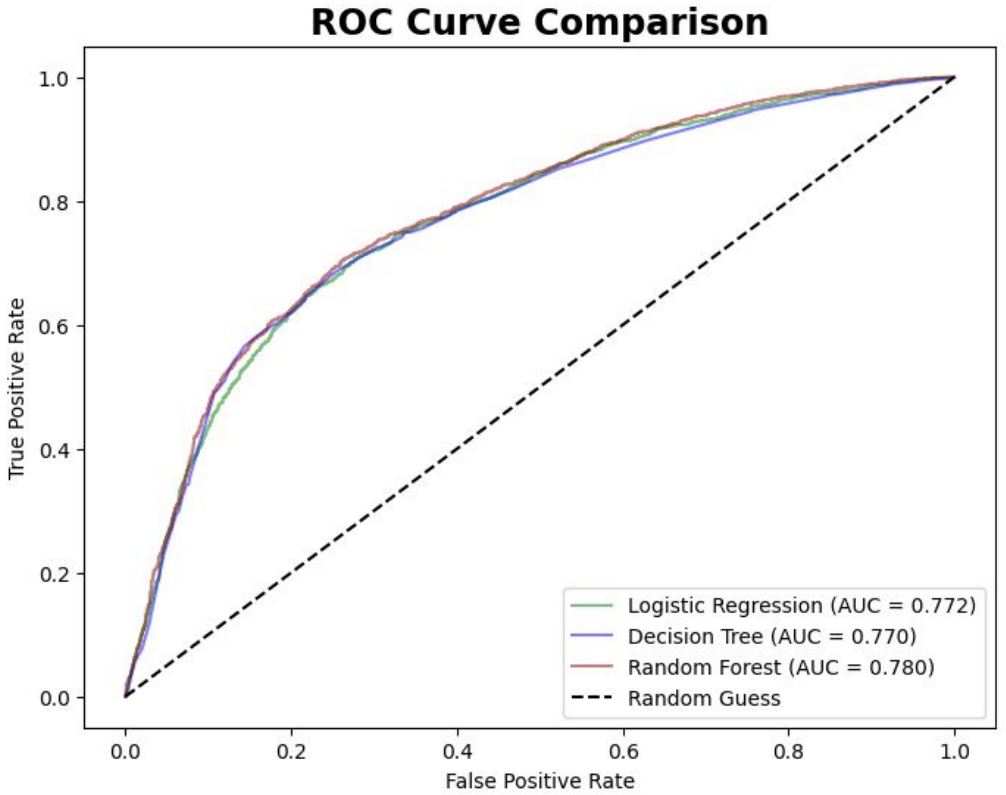
Results & Model Comparison

- Very similar performance scores among all three models
- Logistic regression slightly outperformed other models in terms of F1 score and AUC score



Results & Model Comparison

- ROC curve overlay visualizes model performance



Results & Model Comparison

- Calculated feature importance coefficients for each model
- Systolic blood pressure, age, and cholesterol levels were the 3 most important features in 2 of the 3 models
- Top three features of random forest were systolic blood pressure, age, and diastolic blood pressure
  - Cholesterol was 4th
- Large range of coefficients
  - Only systolic blood pressure for logistic regression and decision tree were above 0.5
  - 70% of all coefficients were below 0.1

Key Takeaways & Summary

- Analyzed the *Cardiovascular Disease dataset* and built models that would predict the presence or absence of CVD based on 11 features and 1 target
- Importance of domain knowledge during data processing to build more accurate models
- All three of our models produced very similar results, with the logistic regression model performing the overall best
- Cholesterol, systolic blood pressure, and age were the top determining factors of CVD

Thank you!