# HEART DISEASE

GBA 6430

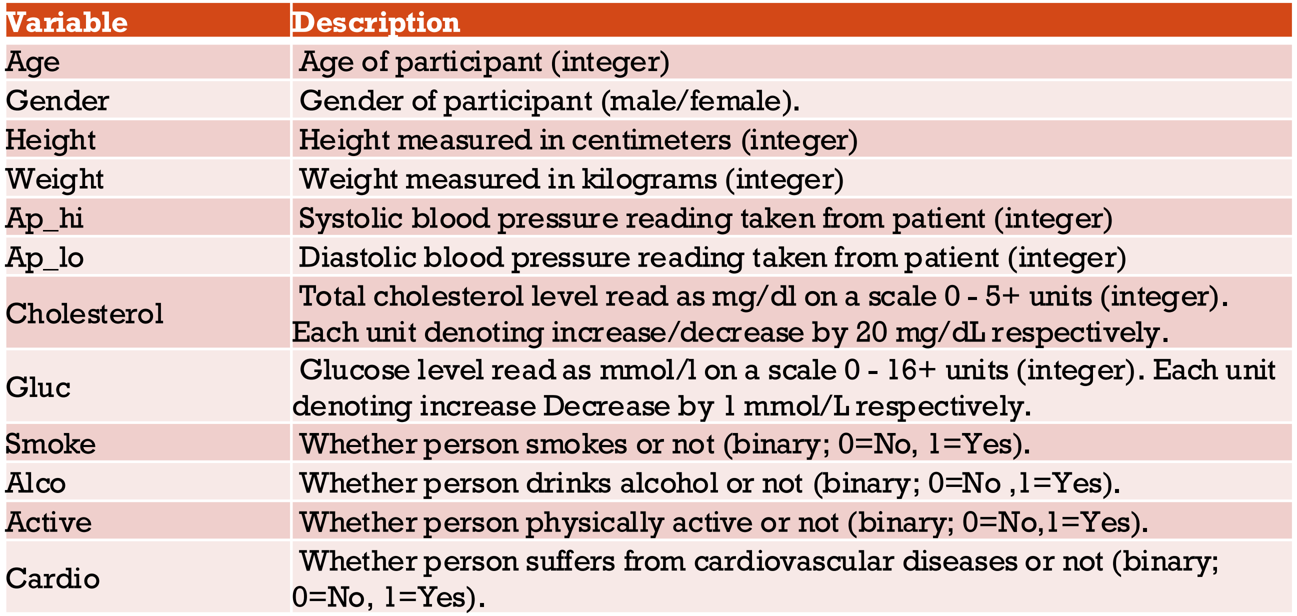
Dr. Salehan

Danica Chin, Gabriela Perez, Kelley Yao

# INTRODUCTION

- Heart disease causes the deaths of 18 million people per year (WHO, 2019)
- Costs the United States approximately $240 billion each year
- This project intends to use anonymous data collected from patients regarding age, weight, blood pressure and glucose measurements to classify potential heart disease
- The expected outcome is to train a classification model to assist in identifying patients who are susceptible to cardiovascular disease earlier in their health history

# DATA COLLECTION



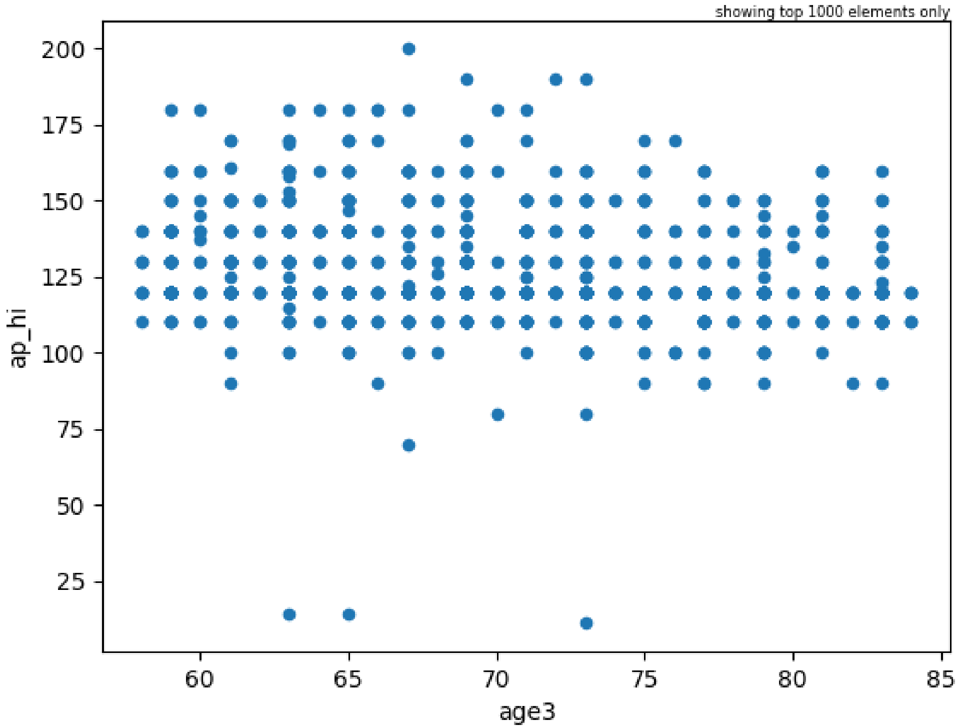
| Variable | Description |
|---|---|
| Age | Age of participant (integer) |
| Gender | Gender of participant (male/female). |
| Height | Height measured in centimeters (integer) |
| Weight | Weight measured in kilograms (integer) |
| Ap_hi | Systolic blood pressure reading taken from patient (integer) |
| Ap_lo | Diastolic blood pressure reading taken from patient (integer) |
| Cholesterol | Total cholesterol level read as mg/dl on a scale 0 - 5+ units (integer). Each unit denoting increase/decrease by 20 mg/dL respectively. |
| Gluc | Glucose level read as mmol/l on a scale 0 - 16+ units (integer). Each unit denoting increase Decrease by 1 mmol/L respectively. |
| Smoke | Whether person smokes or not (binary; 0=No, 1=Yes). |
| Alco | Whether person drinks alcohol or not (binary; 0=No ,1=Yes). |
| Active | Whether person physically active or not (binary; 0=No,1=Yes). |
| Cardio | Whether person suffers from cardiovascular diseases or not (binary; 0=No, 1=Yes). |

- No missing values within the dataset
- No duplicates within the dataset
- Eliminated logical outliers, such as age < 0, weight < 25 kg, height > 240 cm which equaled less than 2% of the dataset

# DESCRIPTIVE ANALYSIS

- The dataset consisted of 70,000 patient records; after dropping logical outliers, analysis was performed on 68,885 records
- The average patient possessed the following attributes:
  - Age: 69 years
  - Weight: 74 kg
  - Height: 164 cm
  - Blood pressure 129/96
  - Non-smoker, female, physically active

# DESCRIPTIVE ANALYSIS



- After performing data visualization, no significant relationship was found between age and diastolic blood pressure
- More significant was correlated relationship between gender/smoking and glucose/cholesterol

# DATA ANALYSIS

Utilized Logistic Regression and Random Forest classification models

Both models were able to perform at 72% and above accuracy rate

Dropped glucose as a predictor variable after fine tuning the model - no significant impact on the dependent variable

# PERFORMANCE MEASURES

| | Before Tuning | | After Tuning | |
|---|---|---|---|---|
| | Logistic Regression | Random Forest | Logistic Regression | Random Forest |
| F-1 score | 0.67 | 0.73 | 0.72 | 0.73 |
| Precision | 0.67 | 0.73 | 0.72 | 0.73 |
| Accuracy | 0.67 | 0.73 | 0.72 | 0.73 |
| Recall | | | 0.72 | 0.73 |

# PERFORMANCE MEASURES: TOP PREDICTORS

Top Predictors of CVD Diagnosis:

- age
- weight
- ap_hi (systolic blood pressure)
- ap_lo (diastolic blood pressure)
- cholesterol ★
- glucose

# SUMMARY

Logistic Regression & Random Forest results indicate that cholesterol, followed by weight and systolic blood pressure were the strongest indicators of CVD – of the variables included within the dataset

Random Forest model performed better than Logistic regression, but only by ~1%

# IMPLICATIONS

- Early prediction for high- risk individuals of cardiovascular disease is vital
  - Enable healthcare providers to intervene early
  - Improving patient outcomes
  - Reducing healthcare costs
  - Enhancing overall quality of care

# LIMITATIONS

- Dataset
- Too few variables. Useful to have
  - Family history
  - Other chronic diseases (co-morbidities)
  - Other vital metrics (LDL & HDL cholesterol, etc.)
- Cholesterol, Alcohol use, physical activity, etc. depicted within a range/binary  instead of actual cholesterol level; would need actual numbers to be helpful in real world

# THANK YOU