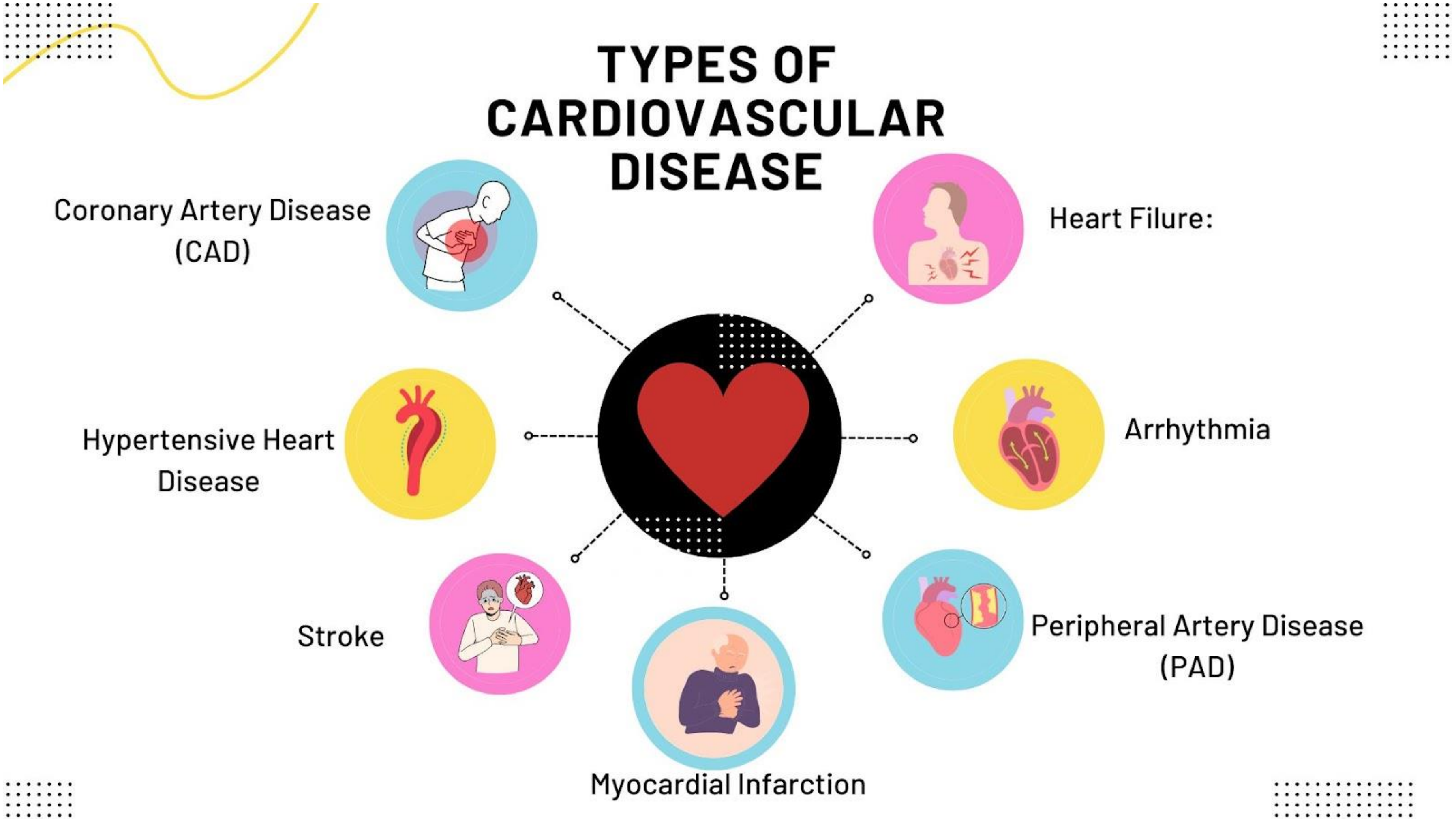
# Studying the Hidden Patterns in Cardiovascular Disease (CVD)

A Data-Driven Exploration of Demographic, Socioeconomic, and Lifestyle Factors Influencing CVD

Presented by:
**Benjamin Luo**

INTRODUCTION

# Cardiovascular Diseases are conditions affecting the heart and blood vessels


TYPES OF CARDIOVASCULAR DISEASE
Coronary Artery Disease (CAD)
Heart Filure:
Hypertensive Heart Disease
Arrhythmia
Stroke
Myocardial Infarction
Peripheral Artery Disease (PAD)

**This analysis will examine:**

**Angina**: chest pain or discomfort caused by reduced blood flow to the heart muscle.

**Coronary Heart Disease** (CHD): the coronary arteries become narrowed or blocked

**Myocardial Infarction** (heart attack): blood flow to the heart muscle is severely blocked

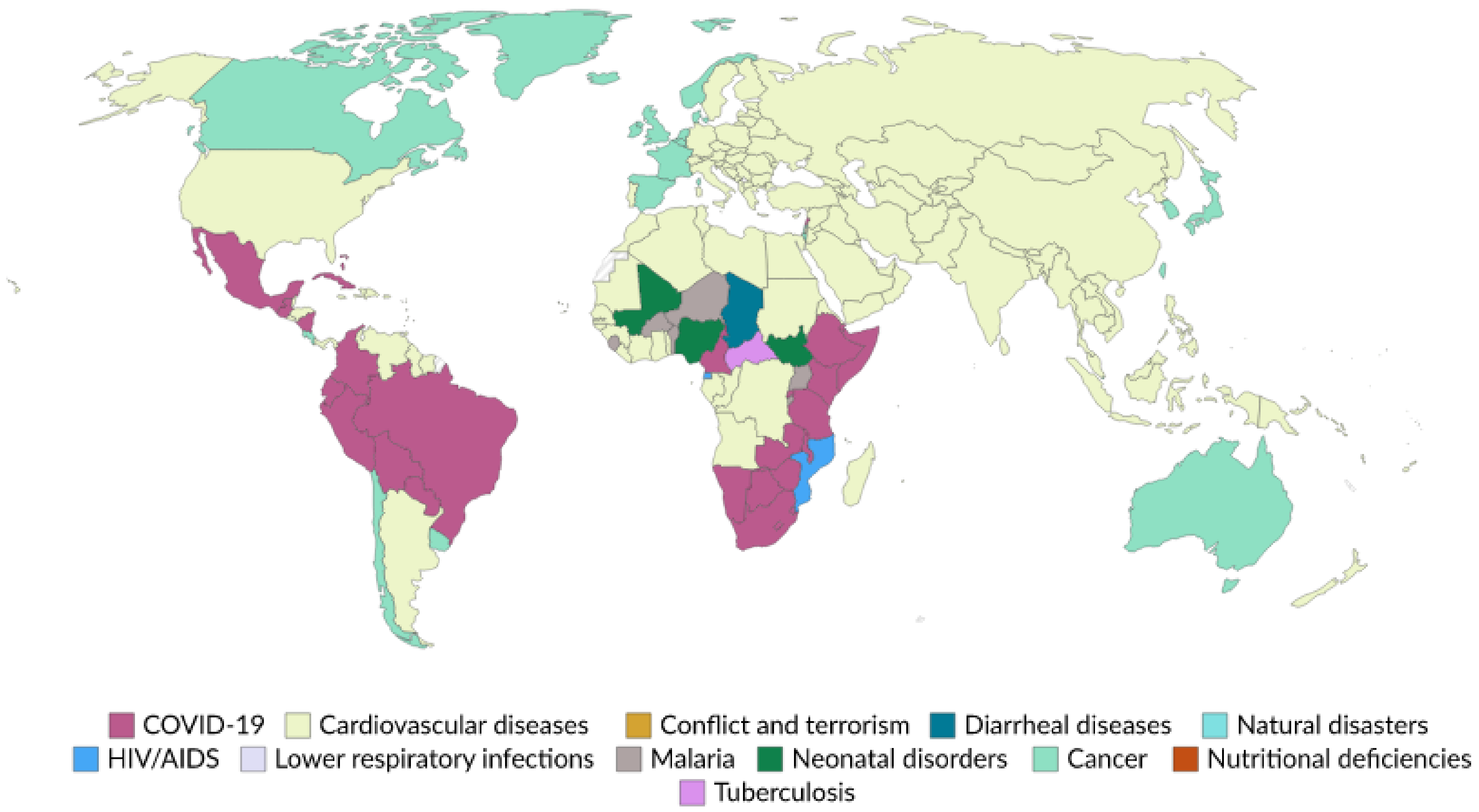
Source: Cardiovascular Disease: Types, Causes & Symptoms. https://www.gargashhospital.com/blog-details/22/cardiovascular-disease-types-causes-and-symptoms. Gargash Hospital

MOTIVATION

# Cardiovascular diseases are the leading cause of death worldwide

## Leading causes of death, 2021

The disease, condition, or injury estimated to cause the most deaths in each country annually.



Our World in Data

**Data source:** IHME, Global Burden of Disease (2024)

OurWorldinData.org/causes-of-death | CC BY

**Note:** Causes of death from different levels from the IHME's disease hierarchy are used in this visualization.

One person dies **every 33 seconds** from cardiovascular disease

In 2022, **702,880 people died** from heart disease (1 in 5 deaths)

Someone has a heart attack **every 40 seconds**

**1 in 5 heart attacks are silent** – the damage is done but the person isn't aware of it

Source: June 2021. Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). World Health Organization (WHO) Heart Disease Facts. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html. US Centers for Disease Control and Prevention (CDC)

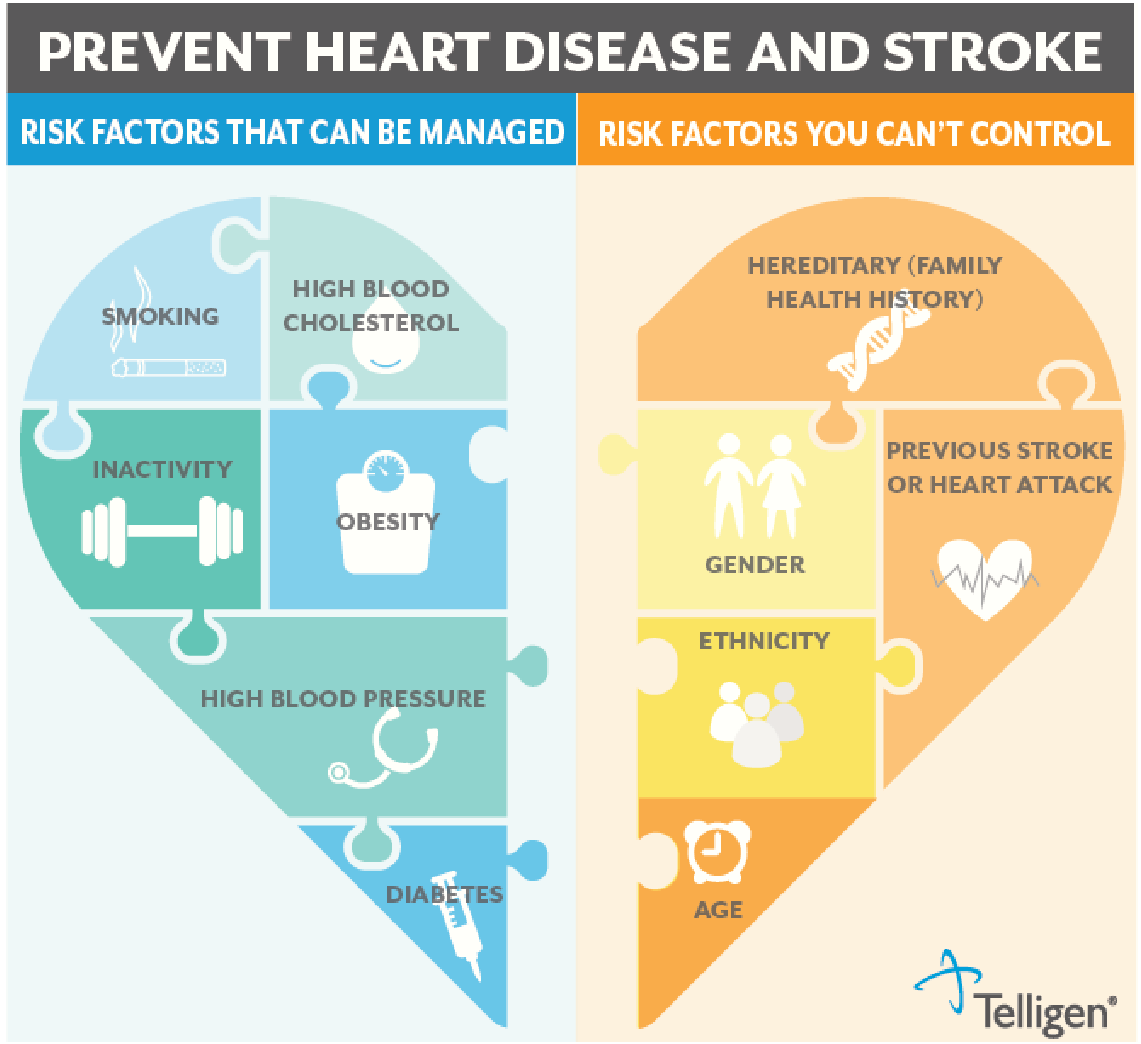
OBJECTIVE

# Exploring the impact of demographics, socioeconomics & lifestyle factors on CVD prevalence

**Demographics**
What is the prevalence of CVD across different age, sex, and ethnic groups?

**Socioeconomics**
Are there significant differences in CVD prevalence based on socioeconomic factors such as income, education, or insurance status?

**Lifestyle**
How do behavioral and lifestyle choices such as smoking, alcohol use, and physical activity influence the prevalence of CVD?


PREVENT HEART DISEASE AND STROKE
RISK FACTORS THAT CAN BE MANAGED
RISK FACTORS YOU CAN'T CONTROL
SMOKING
HIGH BLOOD CHOLESTEROL
INACTIVITY
OBESITY
HIGH BLOOD PRESSURE
DIABETES
HEREDITARY (FAMILY HEALTH HISTORY)
GENDER
PREVIOUS STROKE OR HEART ATTACK
ETHNICITY
AGE
Telligen

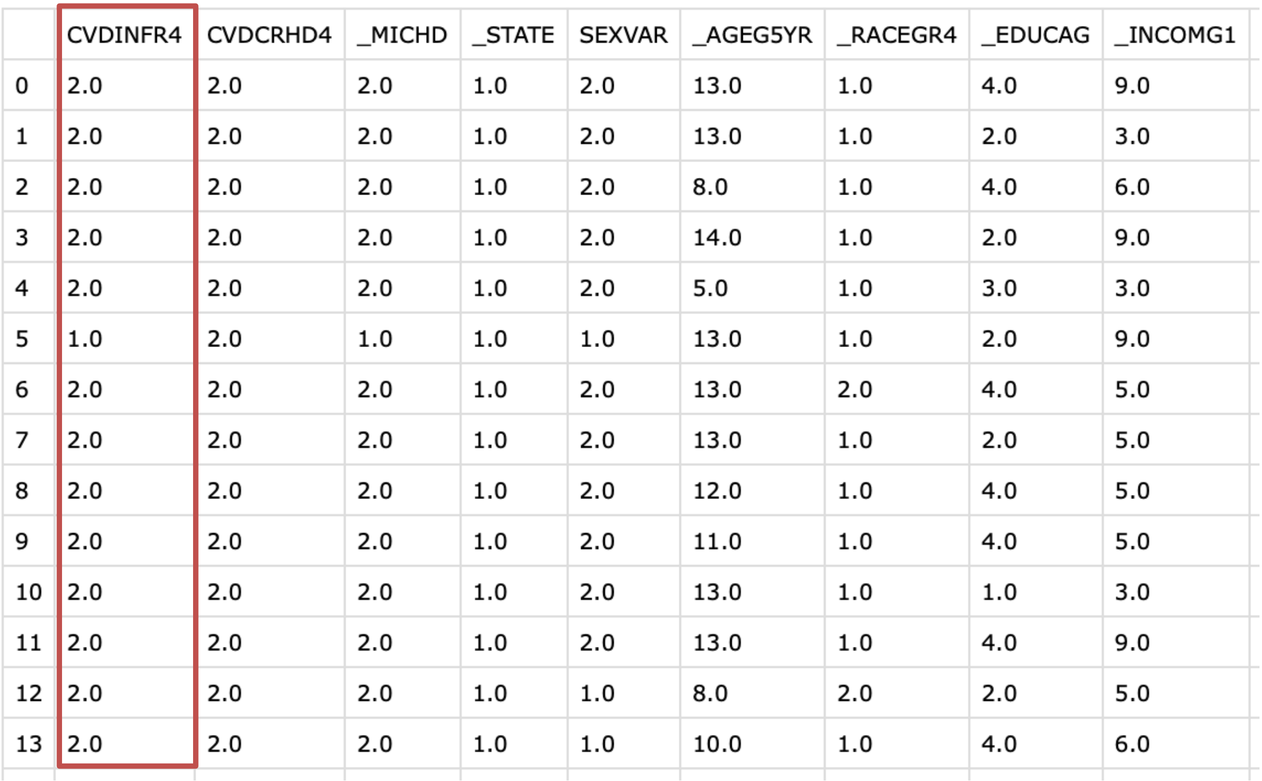
Source: Lower Your Risk for Heart Disease and Stroke (Jan 2018). https://www.telligen.com/lower-your-risk-for-heart-disease-and-stroke. Telligen

DATASET

# The Behavioral Risk Factor Surveillance System (BRFSS) is a CDC dataset tracking US adults' health behaviors and risks

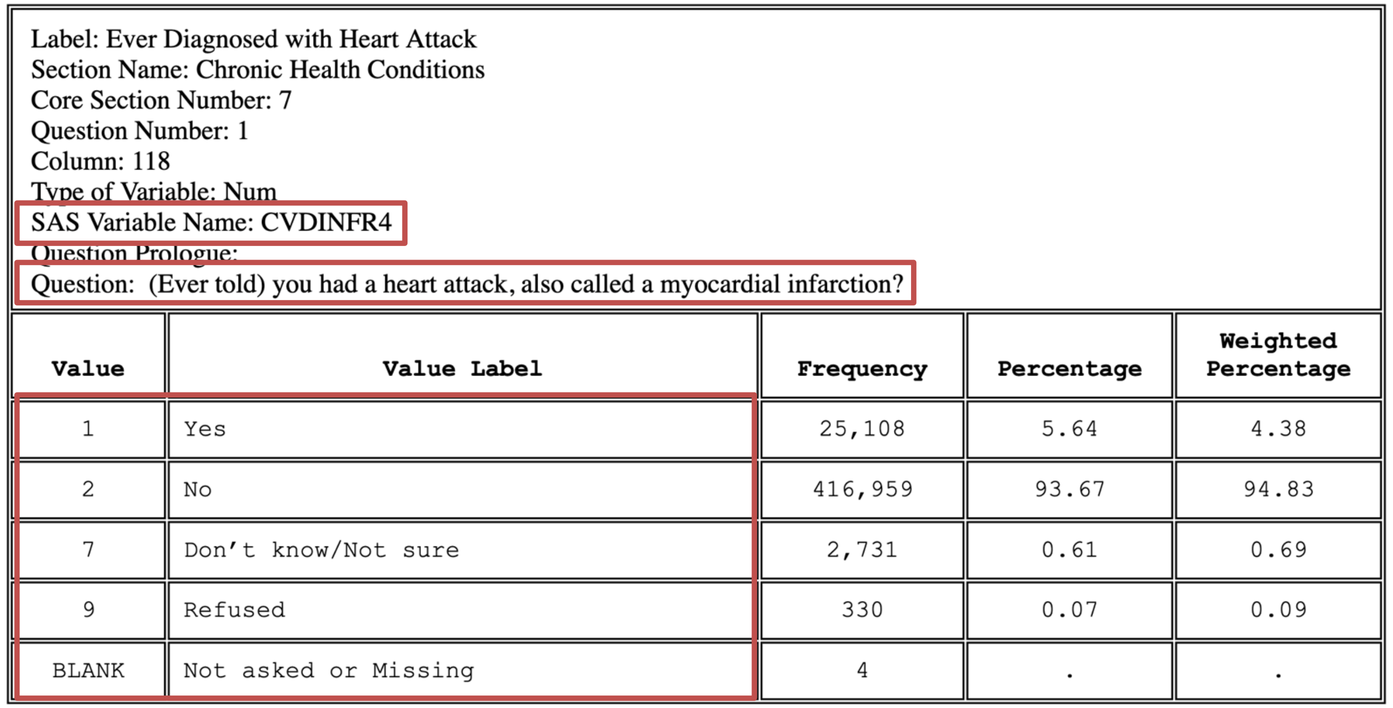
The 2022 dataset contains 445,132 responses and 328 features, of which 36 are studied

## EXCERPT FROM THE BRFSS DATASET



| | CVDINFR4 | CVDCRHD4 | _MICHD | _STATE | SEXVAR | _AGEG5YR | _RACEGR4 | _EDUCAG | _INCOMG1 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2.0 | 2.0 | 2.0 | 1.0 | 2.0 | 13.0 | 1.0 | 4.0 | 9.0 |
| 1 | 2.0 | 2.0 | 2.0 | 1.0 | 2.0 | 13.0 | 1.0 | 2.0 | 3.0 |
| 2 | 2.0 | 2.0 | 2.0 | 1.0 | 2.0 | 8.0 | 1.0 | 4.0 | 6.0 |
| 3 | 2.0 | 2.0 | 2.0 | 1.0 | 2.0 | 14.0 | 1.0 | 2.0 | 9.0 |
| 4 | 2.0 | 2.0 | 2.0 | 1.0 | 2.0 | 5.0 | 1.0 | 3.0 | 3.0 |
| 5 | 1.0 | 2.0 | 1.0 | 1.0 | 1.0 | 13.0 | 1.0 | 2.0 | 9.0 |
| 6 | 2.0 | 2.0 | 2.0 | 1.0 | 2.0 | 13.0 | 2.0 | 4.0 | 5.0 |
| 7 | 2.0 | 2.0 | 2.0 | 1.0 | 2.0 | 13.0 | 1.0 | 2.0 | 5.0 |
| 8 | 2.0 | 2.0 | 2.0 | 1.0 | 2.0 | 12.0 | 1.0 | 4.0 | 5.0 |
| 9 | 2.0 | 2.0 | 2.0 | 1.0 | 2.0 | 11.0 | 1.0 | 4.0 | 5.0 |
| 10 | 2.0 | 2.0 | 2.0 | 1.0 | 2.0 | 13.0 | 1.0 | 1.0 | 3.0 |
| 11 | 2.0 | 2.0 | 2.0 | 1.0 | 2.0 | 13.0 | 1.0 | 4.0 | 9.0 |
| 12 | 2.0 | 2.0 | 2.0 | 1.0 | 1.0 | 8.0 | 2.0 | 2.0 | 5.0 |
| 13 | 2.0 | 2.0 | 2.0 | 1.0 | 1.0 | 10.0 | 1.0 | 4.0 | 6.0 |

## SAMPLE BRFSS QUESTION FROM THE DATA DICTIONARY

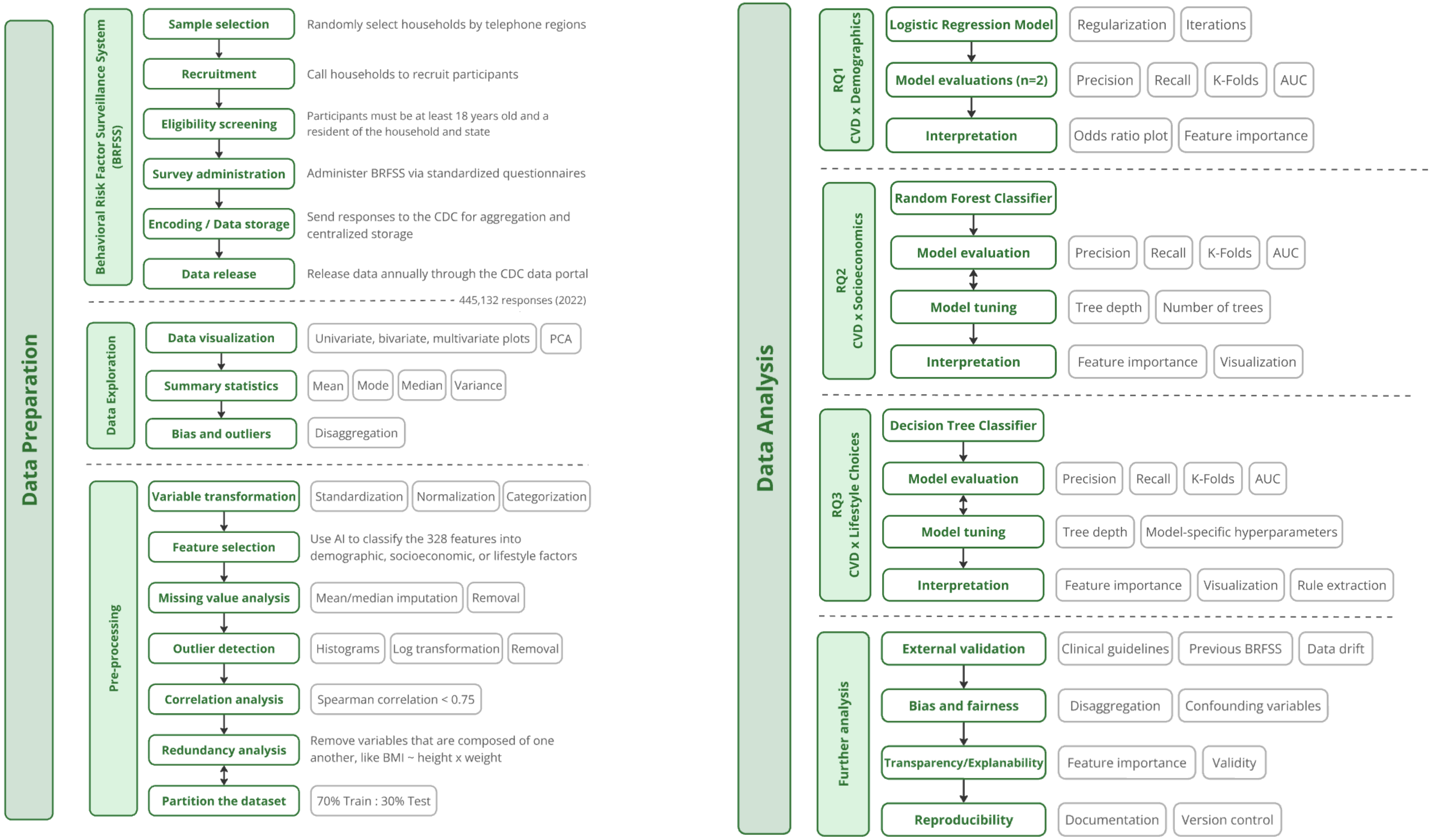


Label: Ever Diagnosed with Heart Attack
Section Name: Chronic Health Conditions
Core Section Number: 7
Question Number: 1
Column: 118
Type of Variable: Num
SAS Variable Name: CVDINFR4
Question Prologue:
Question: (Ever told) you had a heart attack, also called a myocardial infarction?

| Value | Value Label | Frequency | Percentage | Weighted Percentage |
|---|---|---|---|---|
| 1 | Yes | 25,108 | 5.64 | 4.38 |
| 2 | No | 416,959 | 93.67 | 94.83 |
| 7 | Don't know/Not sure | 2,731 | 0.61 | 0.69 |
| 9 | Refused | 330 | 0.07 | 0.09 |
| BLANK | Not asked or Missing | 4 | . | . |

METHODOLOGY

# The data analysis relies on classical approaches that prioritize interpretability for more actionable findings



Data Preparation

Behavioral Risk Factor Surveillance System (BRFSS)

- Sample selection: Randomly select households by telephone regions
- Recruitment: Call households to recruit participants
- Eligibility screening: Participants must be at least 18 years old and a resident of the household and state
- Survey administration: Administer BRFSS via standardized questionnaires
- Encoding / Data storage: Send responses to the CDC for aggregation and centralized storage
- Data release: Release data annually through the CDC data portal

445,132 responses (2022)

Data Exploration

- Data visualization: Univariate, bivariate, multivariate plots | PCA
- Summary statistics: Mean | Mode | Median | Variance
- Bias and outliers: Disaggregation

Pre-processing

- Variable transformation: Standardization | Normalization | Categorization
- Feature selection: Use AI to classify the 328 features into demographic, socioeconomic, or lifestyle factors
- Missing value analysis: Mean/median imputation | Removal
- Outlier detection: Histograms | Log transformation | Removal
- Correlation analysis: Spearman correlation < 0.75
- Redundancy analysis: Remove variables that are composed of one another, like BMI ~ height x weight
- Partition the dataset: 70% Train : 30% Test

Data Analysis

RQ1 CVD x Demographics

- Logistic Regression Model: Regularization | Iterations
- Model evaluations (n=2): Precision | Recall | K-Folds | AUC
- Interpretation: Odds ratio plot | Feature importance

RQ2 CVD x Socioeconomics

- Random Forest Classifier
- Model evaluation: Precision | Recall | K-Folds | AUC
- Model tuning: Tree depth | Number of trees
- Interpretation: Feature importance | Visualization

RQ3 CVD x Lifestyle Choices

- Decision Tree Classifier
- Model evaluation: Precision | Recall | K-Folds | AUC
- Model tuning: Tree depth | Model-specific hyperparameters
- Interpretation: Feature importance | Visualization | Rule extraction

Further analysis

- External validation: Clinical guidelines | Previous BRFSS | Data drift
- Bias and fairness: Disaggregation | Confounding variables
- Transparency/Explainability: Feature importance | Validity
- Reproducibility: Documentation | Version control

**(RQ1)** Identify high-risk groups in healthcare settings

**(RQ2)** Inform policies that promote health equity and access to care

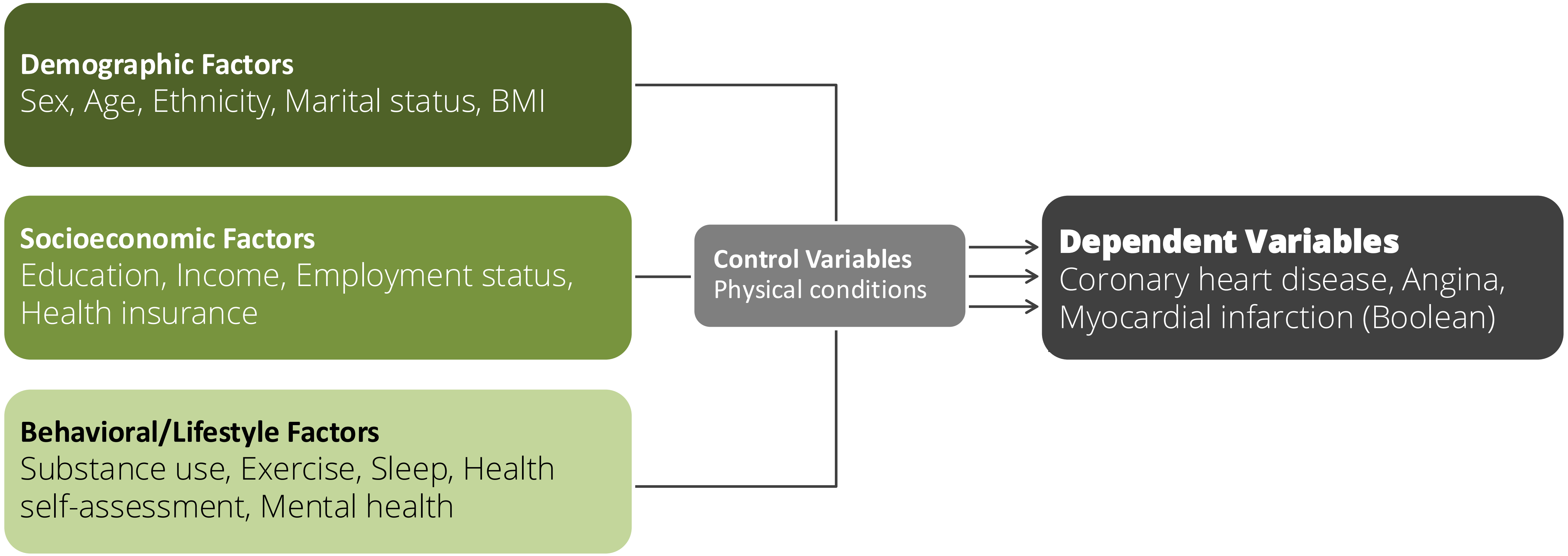
**(RQ3)** Inspire individuals to make healthier lifestyle choices

(Next Steps)

DATASET

# The Behavioral Risk Factor Surveillance System (BRFSS) is a CDC dataset tracking US adults' health behaviors and risks

The 2022 dataset contains 445,132 responses and 328 features, of which 36 are studied



**Demographic Factors**
Sex, Age, Ethnicity, Marital status, BMI

**Socioeconomic Factors**
Education, Income, Employment status, Health insurance

**Behavioral/Lifestyle Factors**
Substance use, Exercise, Sleep, Health self-assessment, Mental health

**Control Variables**
Physical conditions

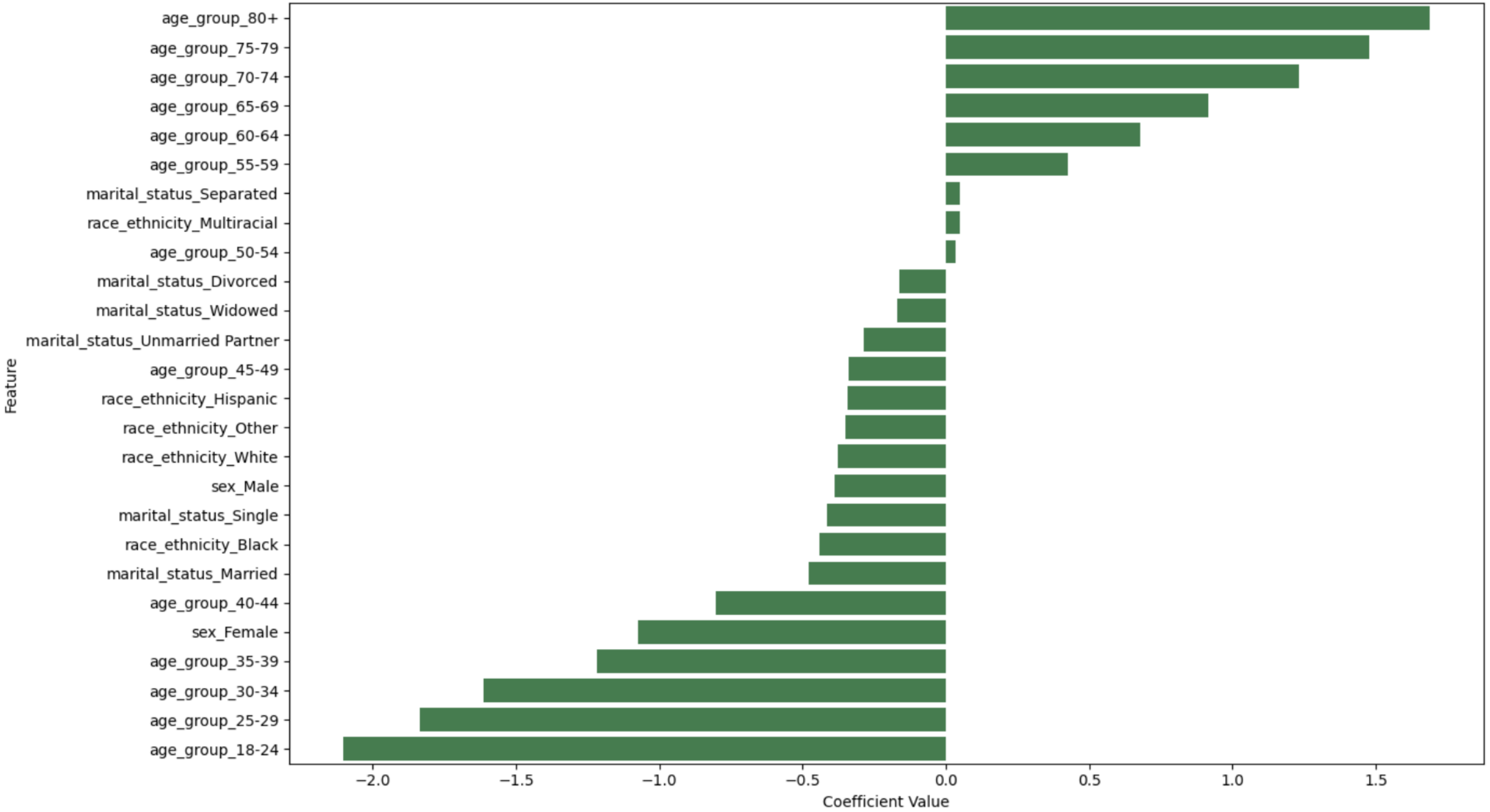
**Dependent Variables**
Coronary heart disease, Angina, Myocardial infarction (Boolean)

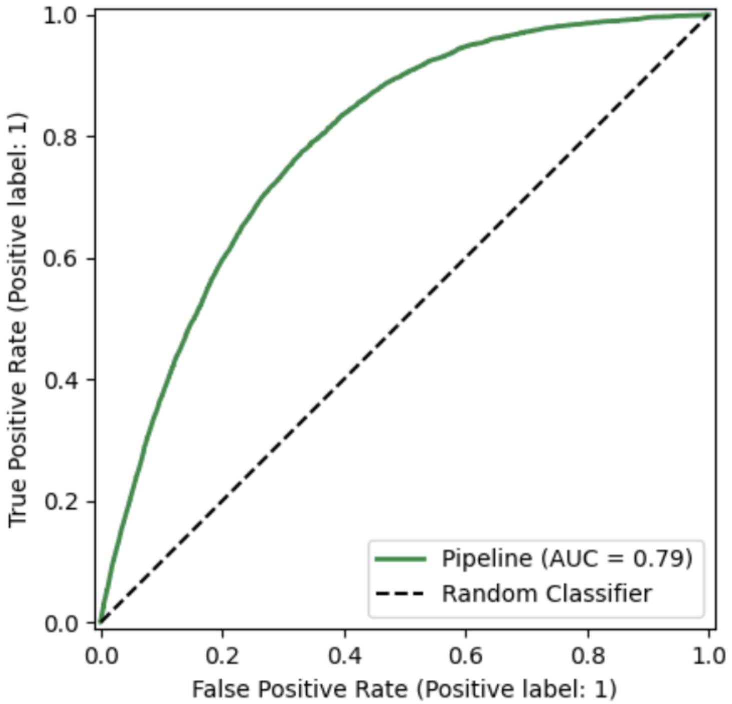
Source: 2022 BRFSS Survey Data and Documentation. cdc.gov/brfss/annual_data/annual_2022.html. US Centers for Disease Control and Prevention (CDC)

ANALYSIS

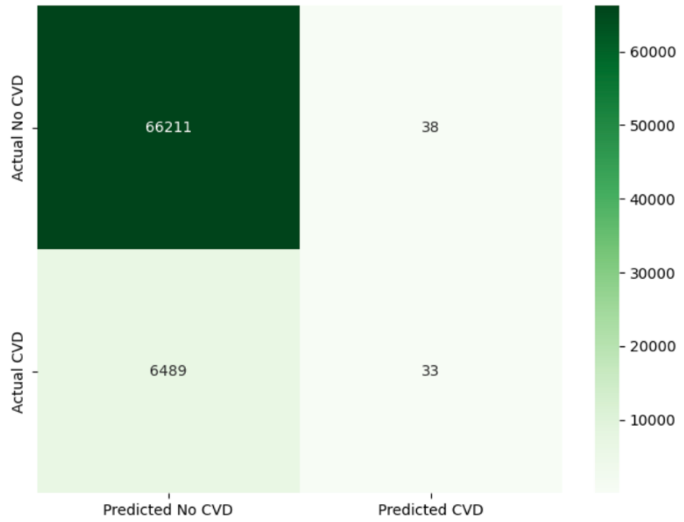
# Of the studied risk factors, the most important demographic factor is age, especially those above the age of 55

LOGISTIC REGRESSION COEFFICIENTS (Feature Importance)





ROC CURVE (AUC = 0.79)

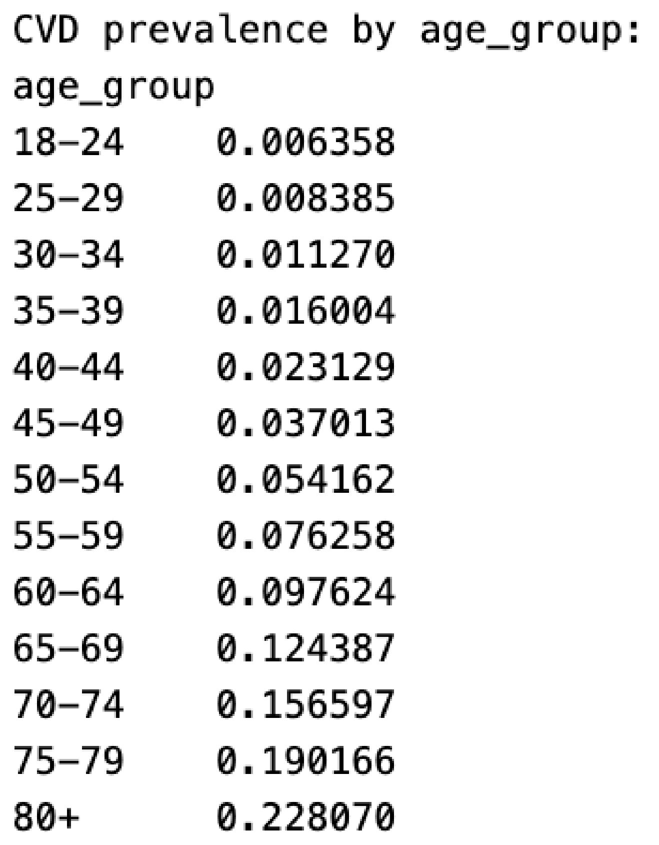


CONFUSION MATRIX

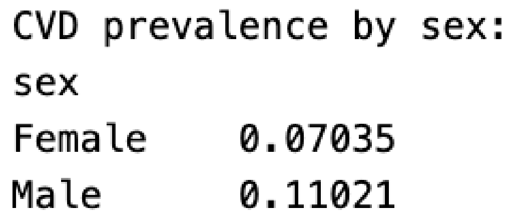
ANALYSIS

# Of the studied risk factors, the most important demographic factor is age, especially those above the age of 55

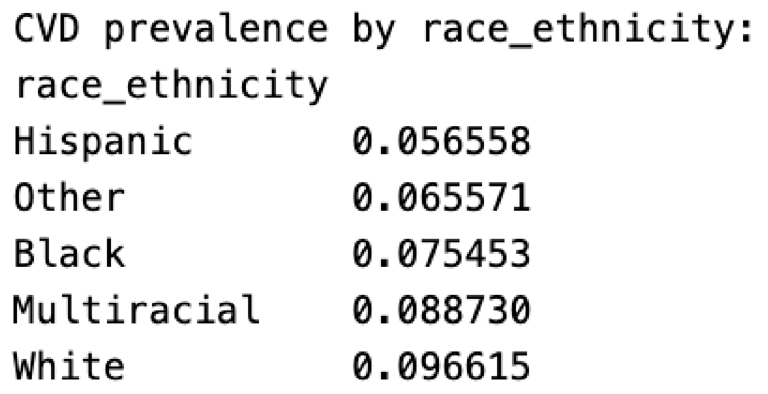
We can also explore prevalence within groups



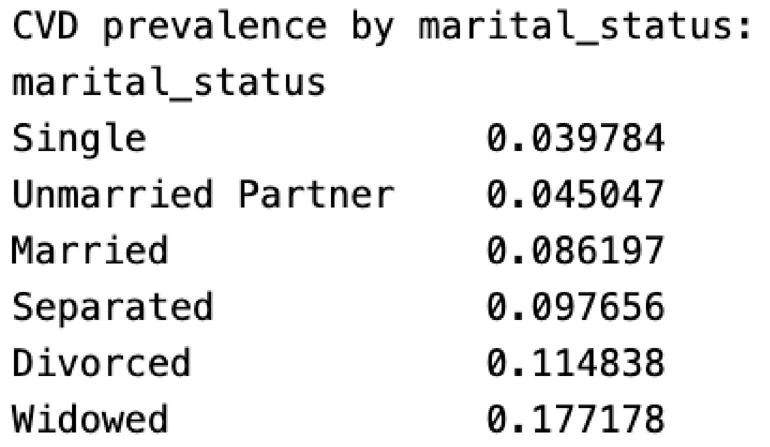
```
CVD prevalence by age_group:
age_group
18-24    0.006358
25-29    0.008385
30-34    0.011270
35-39    0.016004
40-44    0.023129
45-49    0.037013
50-54    0.054162
55-59    0.076258
60-64    0.097624
65-69    0.124387
70-74    0.156597
75-79    0.190166
80+      0.228070
```



```
CVD prevalence by sex:
sex
Female    0.07035
Male      0.11021
```



```
CVD prevalence by race_ethnicity:
race_ethnicity
Hispanic       0.056558
Other          0.065571
Black          0.075453
Multiracial    0.088730
White          0.096615
```



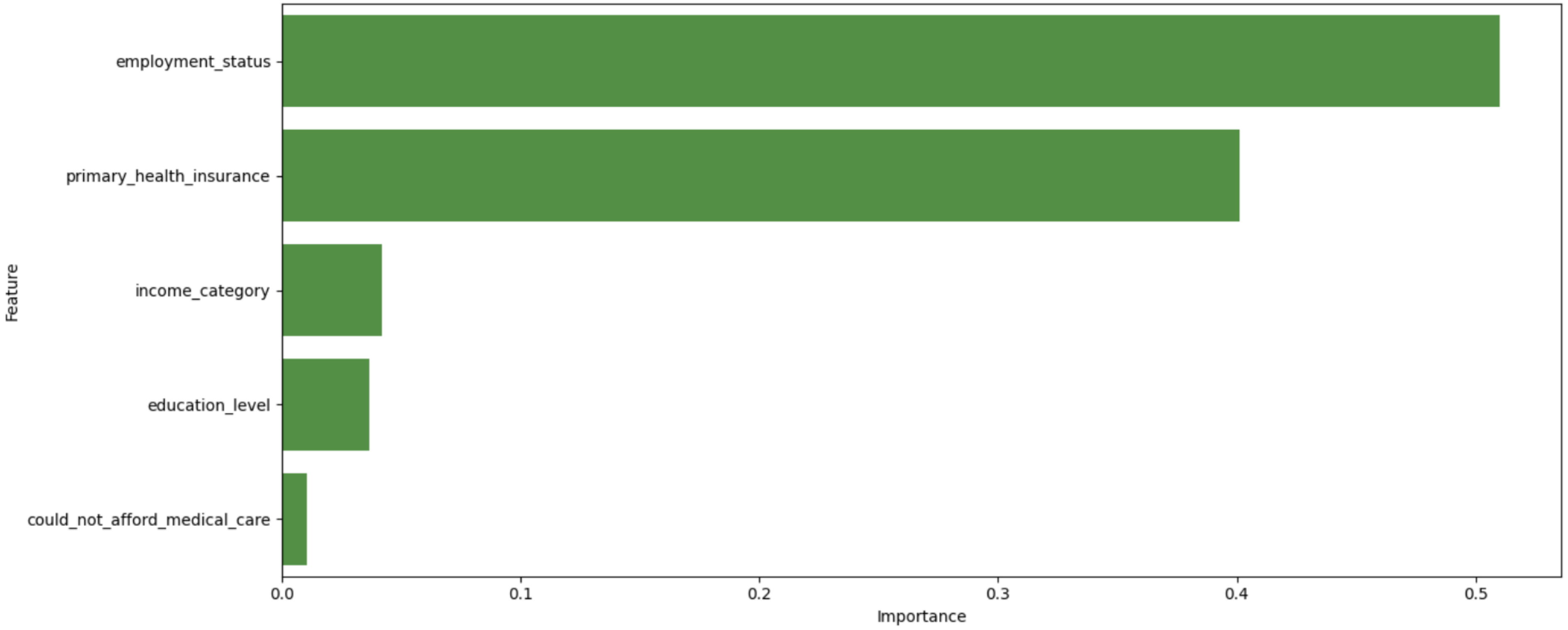
```
CVD prevalence by marital_status:
marital_status
Single               0.039784
Unmarried Partner    0.045047
Married              0.086197
Separated            0.097656
Divorced             0.114838
Widowed              0.177178
```

ANALYSIS

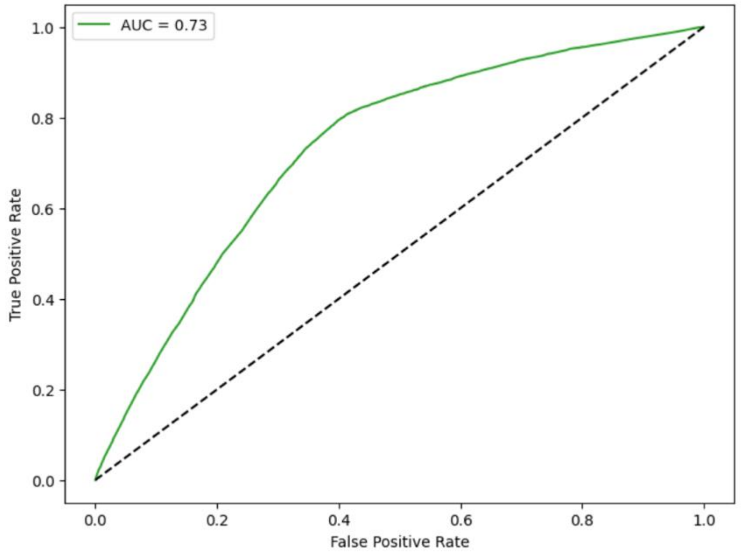
# Of the studied risk factors, the most important socioeconomic factors are employment status and health insurance provider

...but the classifier overfits to the non-CVD class due to an imbalance in the dataset (91% of datapoints are non-CVD)

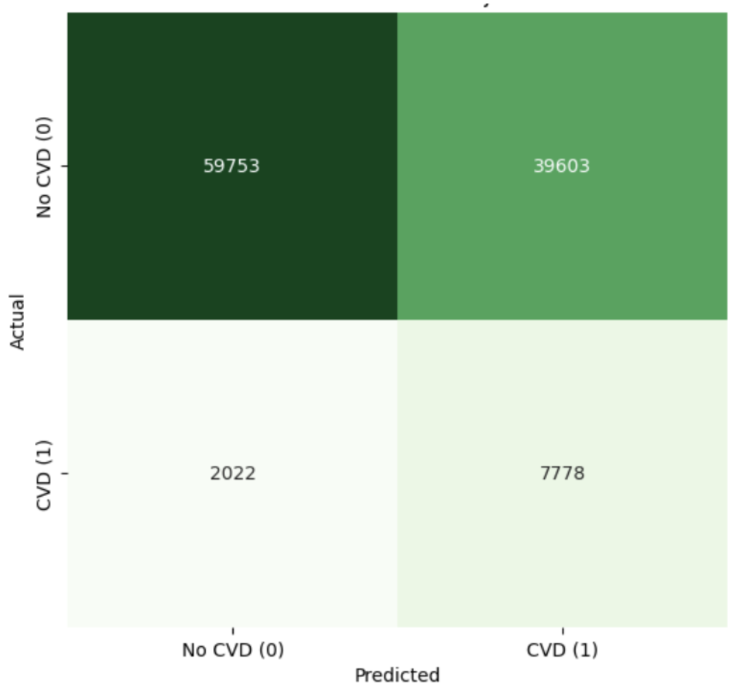
FEATURE IMPORTANCE FROM RANDOM FOREST CLASSIFIER



ROC CURVE (AUC = 0.73)



CONFUSION MATRIX

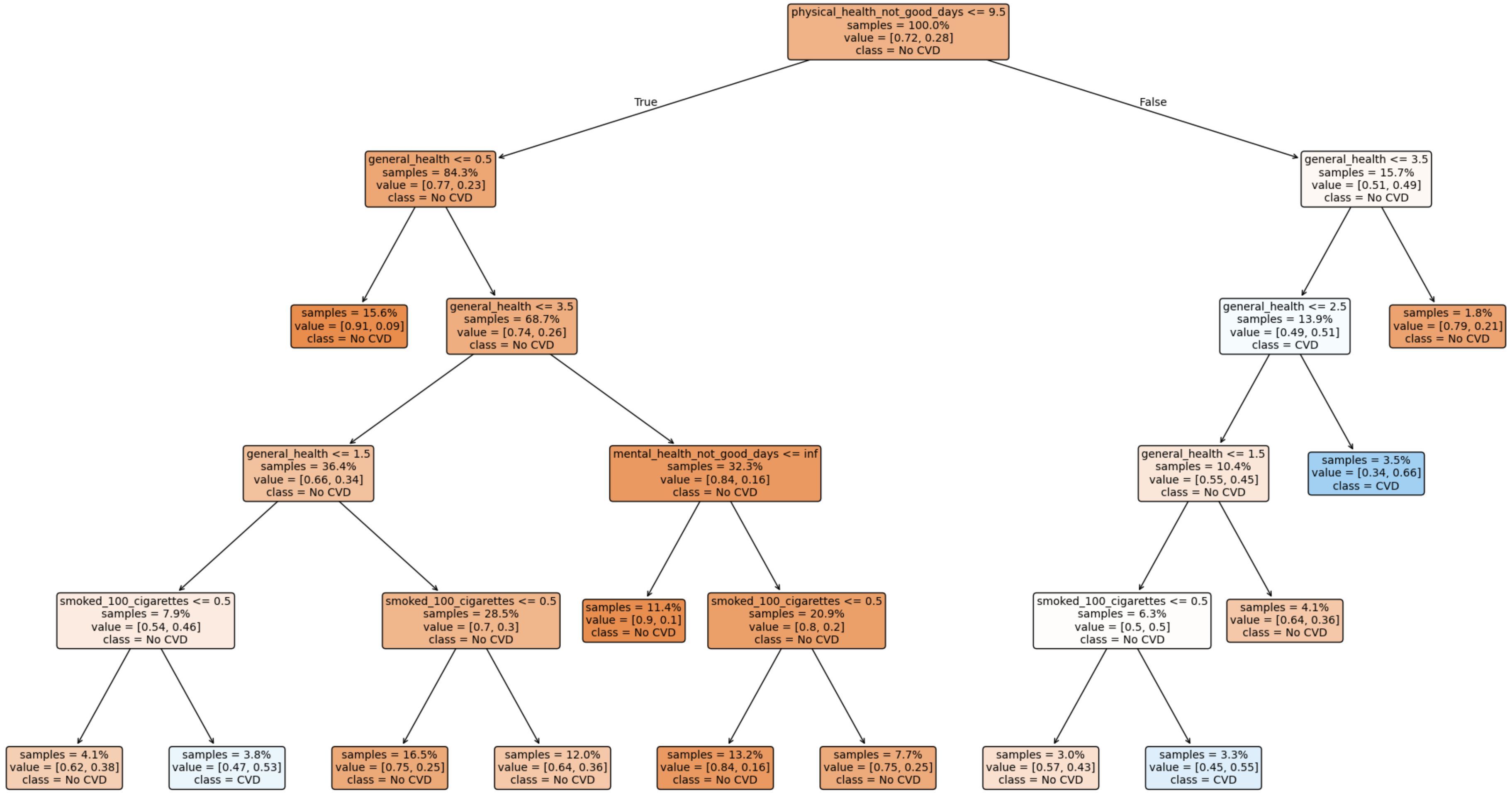


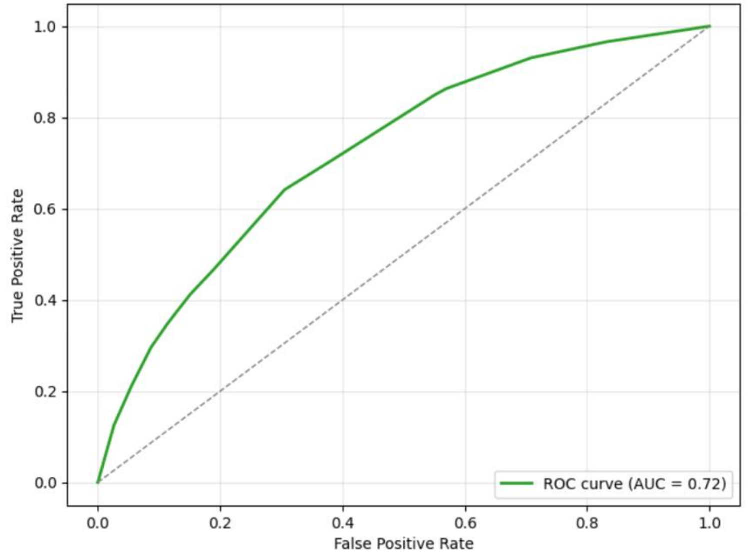
| | No CVD (0) | CVD (1) |
|---|---|---|
| No CVD (0) | 59753 | 39603 |
| CVD (1) | 2022 | 7778 |

ANALYSIS

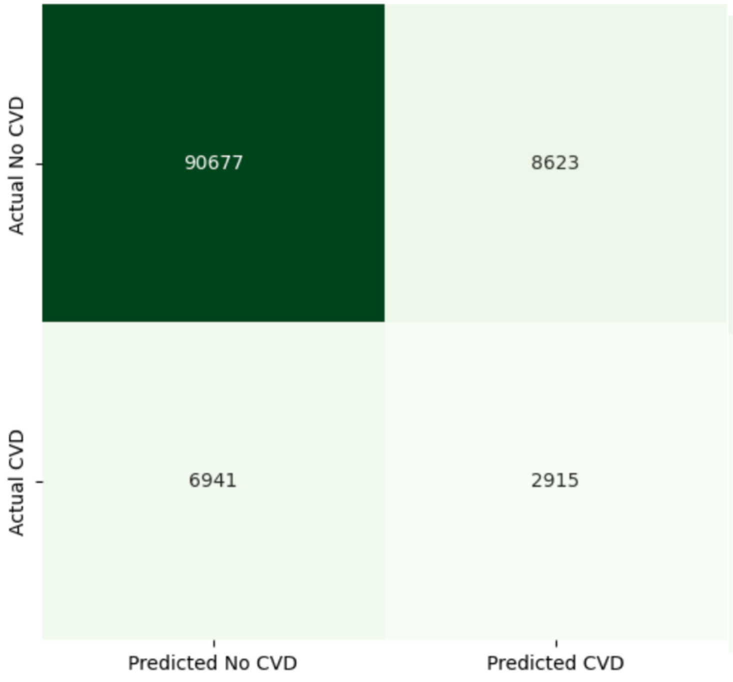
# Of the studied risk factors, the most important behavioral/lifestyle factors are lack of exercise and a low self-assessment of general health

## DECISION TREE FOR CVD RISK (Natural Prevalence ~9%)





ROC CURVE (AUC = 0.72)



|  | Predicted No CVD | Predicted CVD |
|---|---|---|
| Actual No CVD | 90677 | 8623 |
| Actual CVD | 6941 | 2915 |

CONFUSION MATRIX

DISCUSSION

# The findings are consistent with the current clinical knowledge on CVD risk factors

…but further disaggregation is necessary for more granular findings, especially by ethnicity and specific CVDs

| Class | Trait | Risk factor | Possible cause |
|---|---|---|---|
| Demographics | Age | Increasing age | Accumulation of risk factors (ex. diabetes) |
| | Sex | Males | Hormonal differences and lifestyle choices |
| Socioeconomics | Primary health insurer | - | - |
| | Employment | - | - |
| | Income | - | - |
| | Education | - | - |
| Lifestyle choices | Smoking status | Smoker | Damage to blood vessels |
| | Physical activity | Less activity | Exercise helps lower blood pressure |
| | Alcohol consumption | Excess consumption | Increases blood pressure |
| | Stress | Chronic stress | Increasing blood pressure, heart rate, cardiac output |
| | Sleep | Poor sleep quality | Increases hypertension and affects metabolism |

Source: Heart Disease Risk Factors. https://www.cdc.gov/heart-disease/risk-factors/index.html. US Centers for Disease Control and Prevention (CDC).

LIMITATIONS

# While the results are consistent with clinical knowledge, the BRFSS dataset may have limitations, especially for analyzing CVD

### Response Bias

16% of respondents did not disclose their income

### Indirect "Proxy” Measures

Lack of biomarkers (ex. heart rate, blood pressure)

### Subjective Measures

“Rate your general health on a scale from 1-10”

### Imbalanced Dataset

6%-9% reported having a heart attack, CHD, or angina

### CVD Categorization

(CHD + Heart Attack) and (CHD + Angina) were paired together

### Technical Survey Questions

“Myocardial Infarction”, “Angina”, “Coronary Heart Disease”

### Lack of Generalizability

The survey is only administered in the USA

### Lack of Validation

Self-reported diagnoses were not validated through medical records

# Key Points

| Class | Trait | Risk factor |
|---|---|---|
| Demographics | Age | Increasing age |
| | Sex | Males |
| Socioeconomics | Employment | - |
| | Primary health insurer | - |
| | Income | - |
| | Education | - |
| Lifestyle choices | Smoking status | Smoker |
| | Physical activity | Less activity |
| | Alcohol consumption | Excess consumption |
| | Stress | Chronic stress |
| | Sleep | Poor sleep quality |