Predicting the features of heart disease using Cardiovascular Dataset

Presenters:

- Susmi Sharma
- Pratibha Pokharel
- Christina Liam

Introduction

Background

- What is heart disease?
  - The term “heart disease” refers to several types of heart conditions.
  - Coronary artery disease (CAD) affects blood flow to the heart which can cause a heart attack (CDC).

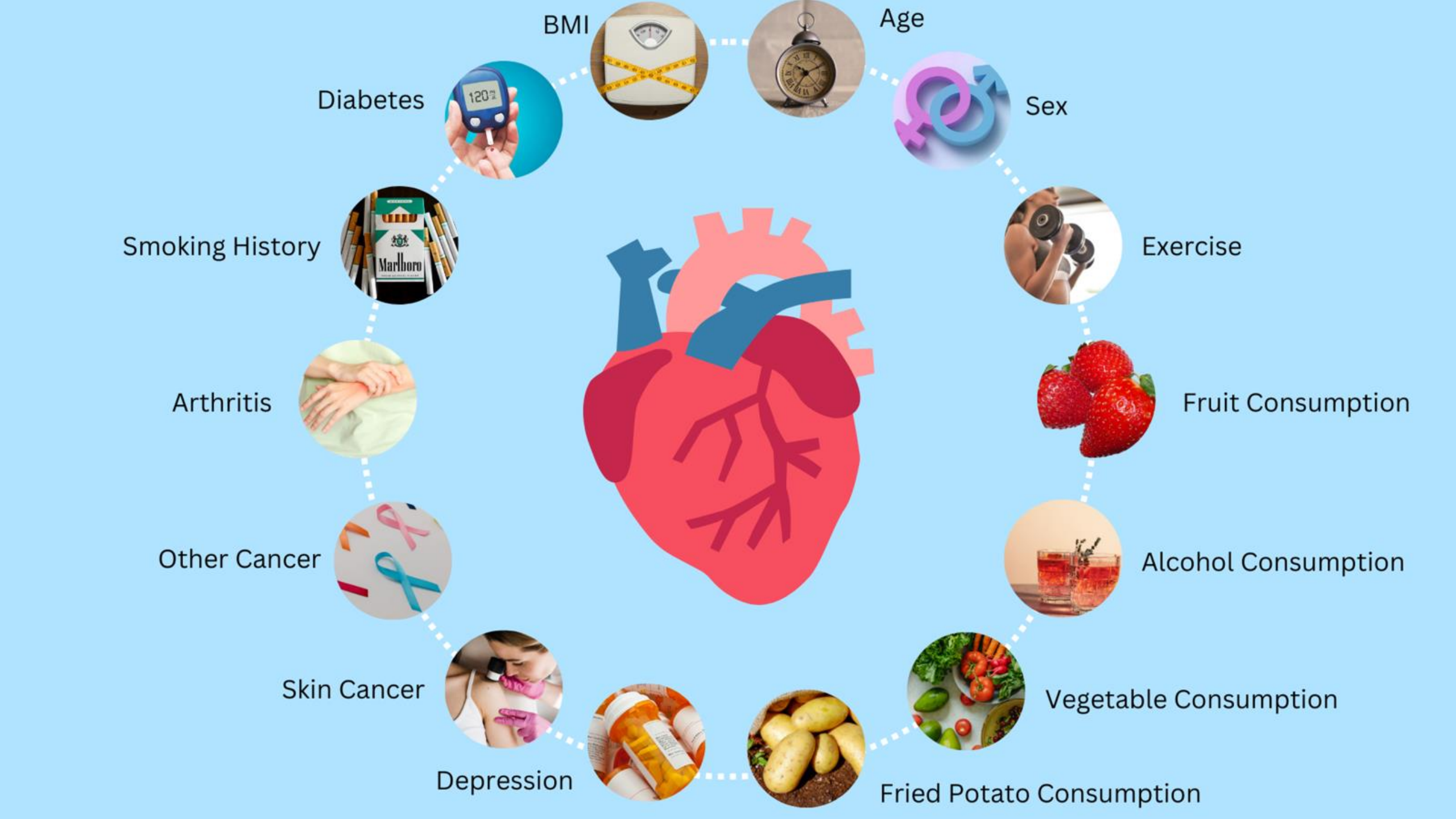
- How many people are globally impacted by it every year?
  - CVD is the leading cause of death globally, taking an estimated 17.9 million lives each year (WHO).

- What are the consequences of heart diseases to one’s heart and their family?
  - If you have a family health history of heart disease, you are more likely to develop heart disease yourself.

Background

- We still do not have a cohesive theory on what causes or the risks factors of heart disease.
  - Research conducted by University of Sydney and Heart Research Australia has found that there is an increasing proportion of heart attack patients without any standard risk factors (cholesterol).

- What has been done and what still needs to be analyzed?
  - In 2023 advances in technology to restore blood flow to blocked and narrowed arteries, potentially preventing death and disability for a wide range of patients, included those with severe illness (American Heart Association).


BMI
Age
Diabetes
Sex
Smoking History
Exercise
Arthritis
Fruit Consumption
Other Cancer
Alcohol Consumption
Skin Cancer
Vegetable Consumption
Depression
Fried Potato Consumption

Literature Review: Variables

Age & Heart Disease

- Age is considered one of the most significant non-modifiable risk factors for heart disease (Benjamin, 2019).

BMI & Heart Disease

- Several studies have demonstrated a positive association between these two variables.

Fruit Consumption & Heart Disease

- Regular consumption of fruits contribute to a lower risk of heart disease (Wang, 2014).

Vegetable Consumption & Heart Disease

- Green vegetables are linked to improved cardiovascular health (Boeing, 2012).

Fried Potato Consumption & Heart Disease

- High consumption, especially in the form of fries and potato chips, has been associated with an increased risk of heart disease.

Alcohol Consumption & Heart Disease

- Excessive consumption can increase the risk of heart disease through various mechanisms such as elevated blood pressure (Ronksley, 2011).

Diabetes & Heart Disease

- This relationship is a major risk factor but is multifactorial (Sarwar, 2010).

General Health & Heart Disease

- Overall general health plays a crucial role in the prevention of heart disease.

Cardiovascular Kaggle Dataset

- Originates from a telephone-based survey conducted in 2021 across 47 states.
- Behavioral Risk Factor Surveillance System (BRFSS) administered by the Centers for Disease Control and Prevention.
- Questionnaire, interview procedures, and the data itself, please see the BRFSS 2021 Overview.
- From 304 features from 438,693 participants, Kaggle condensed to 19 features and 308,854 participants.

Past Approaches

- Included twelve continuous features and seven categorical, with several binary predictors.
- Given the large number of binary predictors and the substantial volume of data, logistic regression and random forest model were the most suitable analytical tool.
- Logistic regression was previously utilized on the dataset (Lupague et al, 2023; Kaggle).
- **Nobody had analyzed this data using random forest in R in an elaborative manner.**

Aims of our Projects

Public health driven

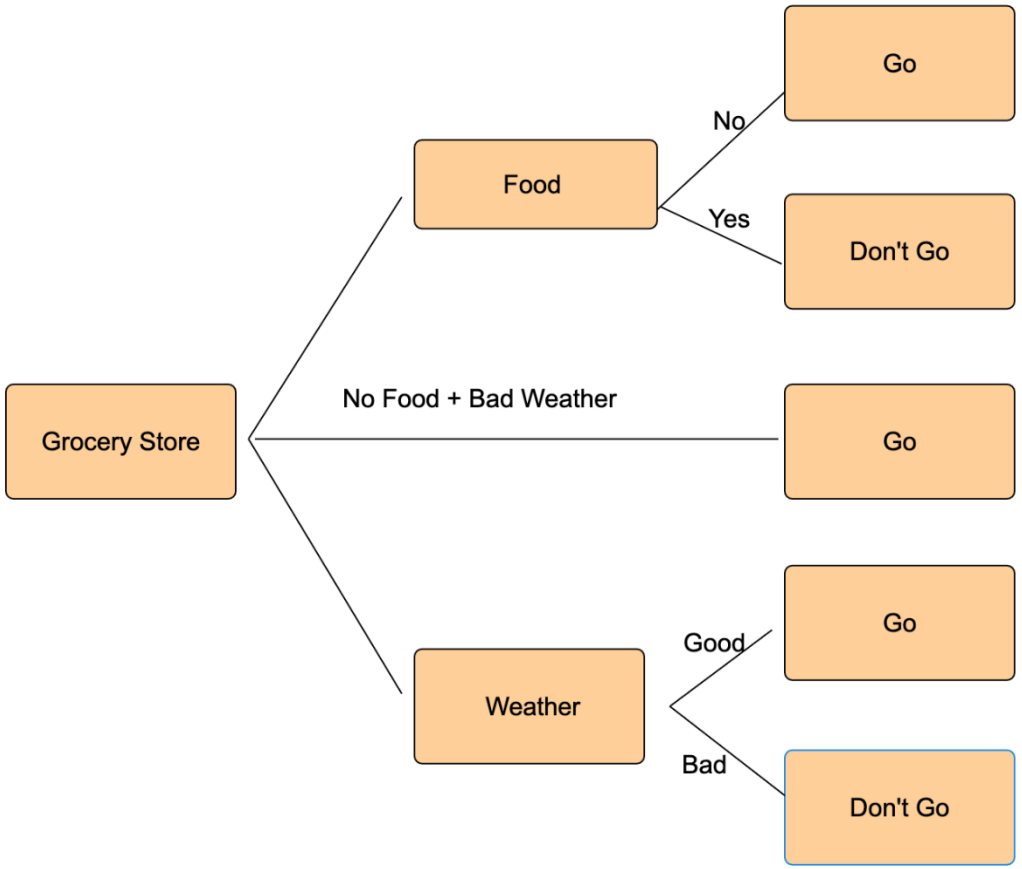
- What are the physical and life-styles characteristics of individuals with and without heart disease?

Computationally driven

- Whether we should handpick the features that seem theoretically relevant or use all the features that are given to us, while investigating health-care related research questions?
- To answer this question, we will analyze the data two different ways:
  - Random forest with a few selected features
  - Random forest with all the features

Random Forest model

- Advanced extension of decision tree modeling.
- mirrors systematic human thought processes in sequential steps
- Unlike a single decision tree, random forest utilizes numerous trees, each analyzing a variable set of decision pathways.
- Reduces overfitting
- Improves the generalizability of the model



Method

Data

- CVD data —no missing data.
- Included a total of 308854 participants
- 19 features — the mental and physical health, dietary habits, and lifestyle choices of the patients.
- 8% (n = 24, 971) of the total participants in the telephone — having heart diseases.

Data Pre-processing

- The age variable was pre-processed the following manner.
  - Diabetes features were also removed.

	Age_Category	Heart_Disease	Age_min	Age_max	AverageAge	Group
1	70-74	Yes	70	74	72	70 years
2	75-79	Yes	75	79	77	70 years
3	60-64	Yes	60	64	62	60 years
4	75-79	Yes	75	79	77	70 years
5	75-79	Yes	75	79	77	70 years
6	75-79	Yes	75	79	77	70 years
7	65-69	Yes	65	69	67	60 years
8	70-74	Yes	70	74	72	70 years
9	75-79	Yes	75	79	77	70 years
10	80+	Yes	80	80	80	80 years
11	75-79	Yes	75	79	77	70 years
12	75-79	Yes	75	79	77	70 years
13	70-74	Yes	70	74	72	70 years
14	75-79	Yes	75	79	77	70 years

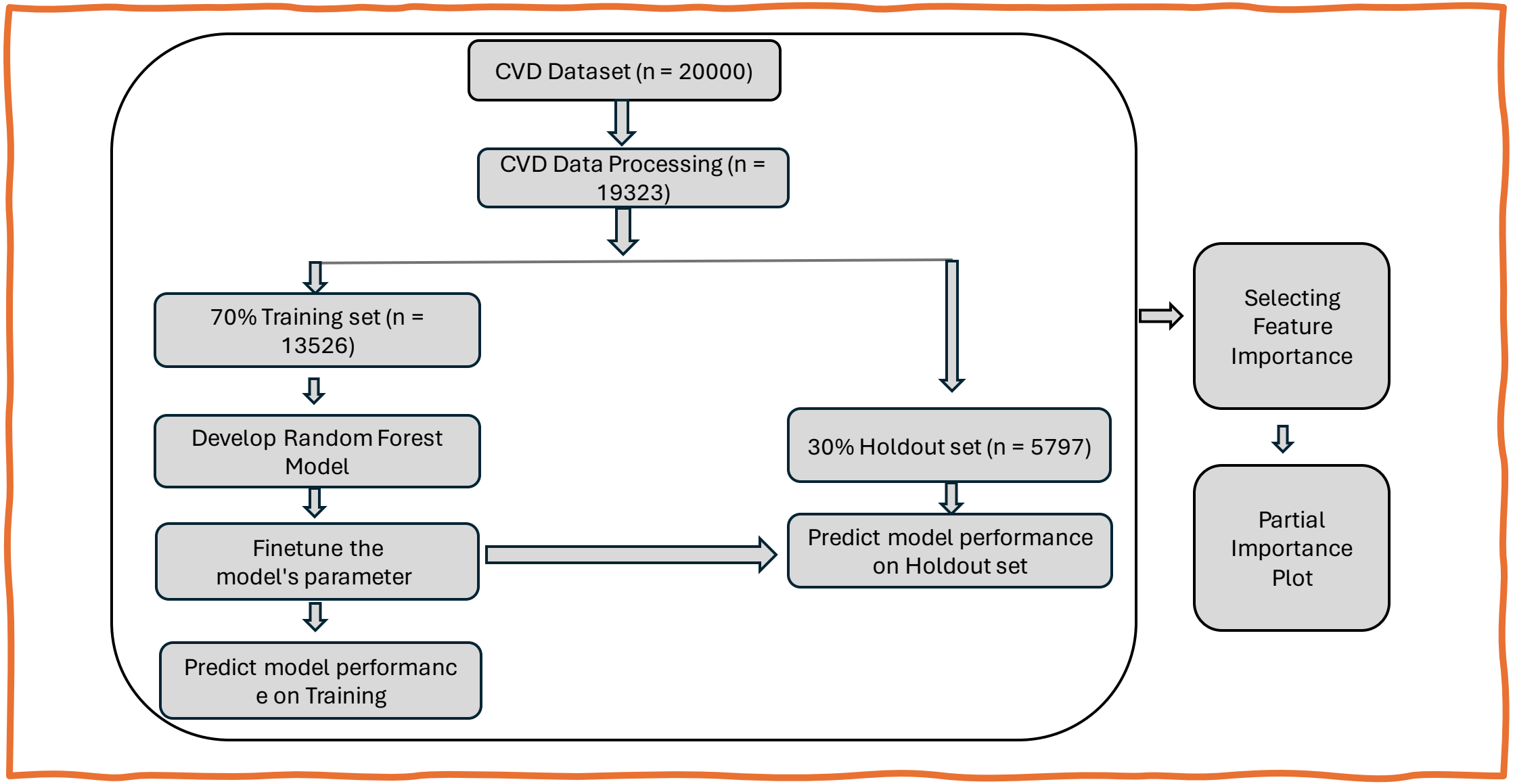
Data Pre-processing

- Reduced the dataset to 20,000 randomly sampled data points —imited computational resources.
  - Excluded 677 observations that fell outside of the "yes" and "no" categories for diabetes.
  - Total of 19,323 observations, comprising
    - 9,649 participants with heart diseases
    - and 9,674 participants without heart diseases.

Model 1: Random Forest with 7 features
Model 2: Random Forest with all Features

Random Forest Model			
Model	**Features**	**Input**	**Output**
1	***Handpicked***	Age, diabetes, fruit consumption, vegetables consumption, fried potatoes consumption, alcohol consumption, and BMI	Heart Disease (Yes/No)
2	***All features***	Age, diabetes, fruit consumption, vegetables consumption, fried potatoes consumption, alcohol, BMI, depression, sex, diabetes, height, weight, general health, exercise, skin cancer, smoking history, checkup, and arthritis.	Heart Disease (Yes/No)

Flow-Chart


CVD Dataset (n = 20000)
CVD Data Processing (n = 19323)
70% Training set (n = 13526)
Develop Random Forest Model
Finetune the model's parameter
Predict model performanc e on Training
30% Holdout set (n = 5797)
Predict model performance on Holdout set
Selecting Feature Importance
Partial Importance Plot

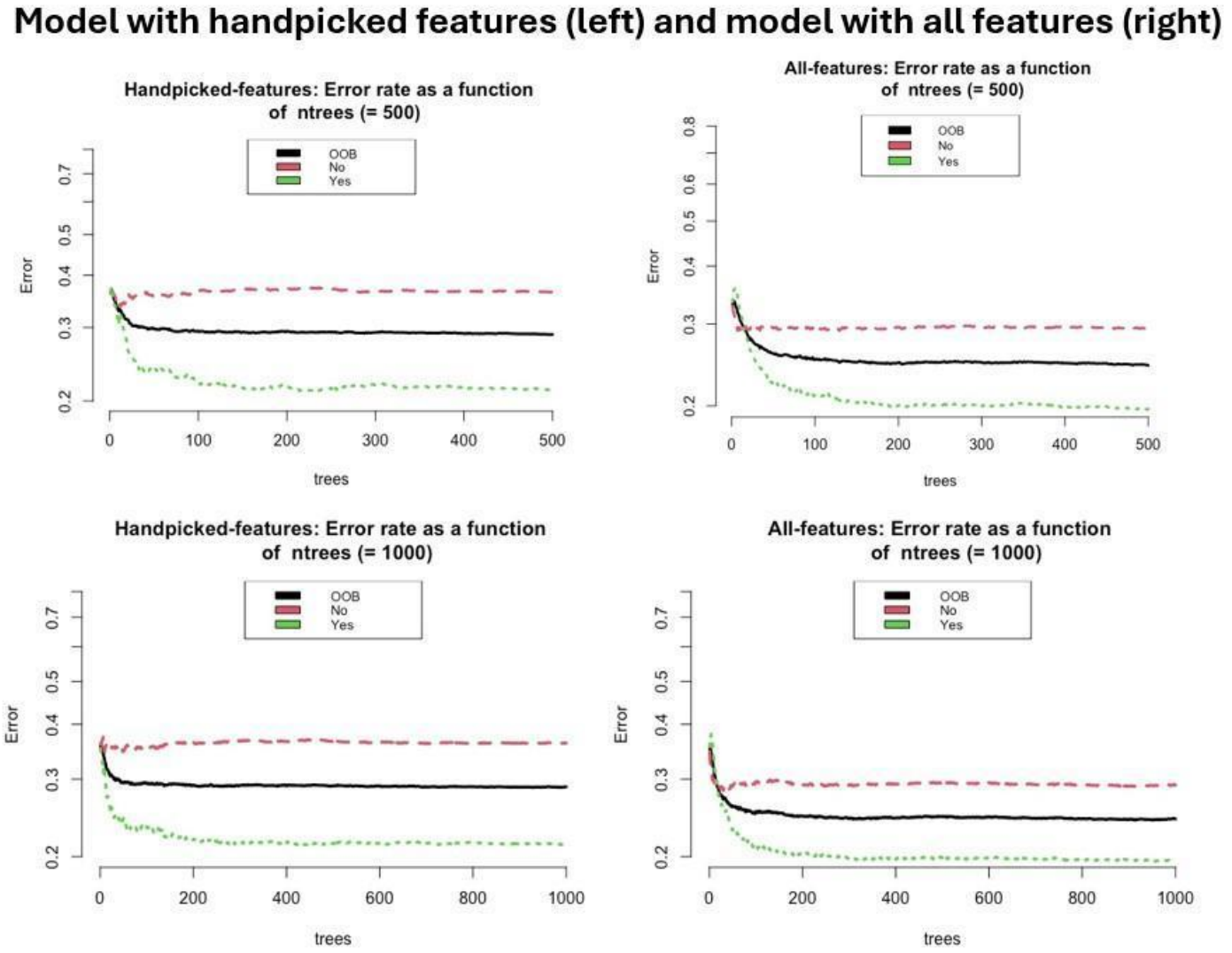
Results

Random forest algorithm to the training data set 1 and 2 and obtained model 1 and model 2 respectively

- Optimized two model parameters.
  - Number of decision trees (ntree)
  - Number of splits at every node of the tree (m)

Model error as a function of number of tree

- 500 decision trees were enough to get a stable out of bag error estimate (OOB)
- split number of 2 —the lowest OOB for both models.



Confusion matrix

Model with 8 features

Predicted Observation	True Observation				
	Training Set			Testing Set	
		Heart Healthy	Heart Disease	Heart Healthy	Heart Disease
	Heart Healthy	5805	413	1851	610
	Heart Disease	943	6365	1075	2261

OOB rate : 28.87%

Testing Set

Specificity = 78.75% ,

Sensitivity = 63.26%,

Model with All features

Predicted Observation	True Observation				
	Training Set			Testing Set	
		Heart Healthy	Heart Disease	Heart Healthy	Heart Disease
	Heart Healthy	6091	384	2022	541
	Heart Disease	657	6394	904	2330

OOB rate: 24.43%

Testing Set

Specificity = 81.16%

Sensitivity = 69.90%

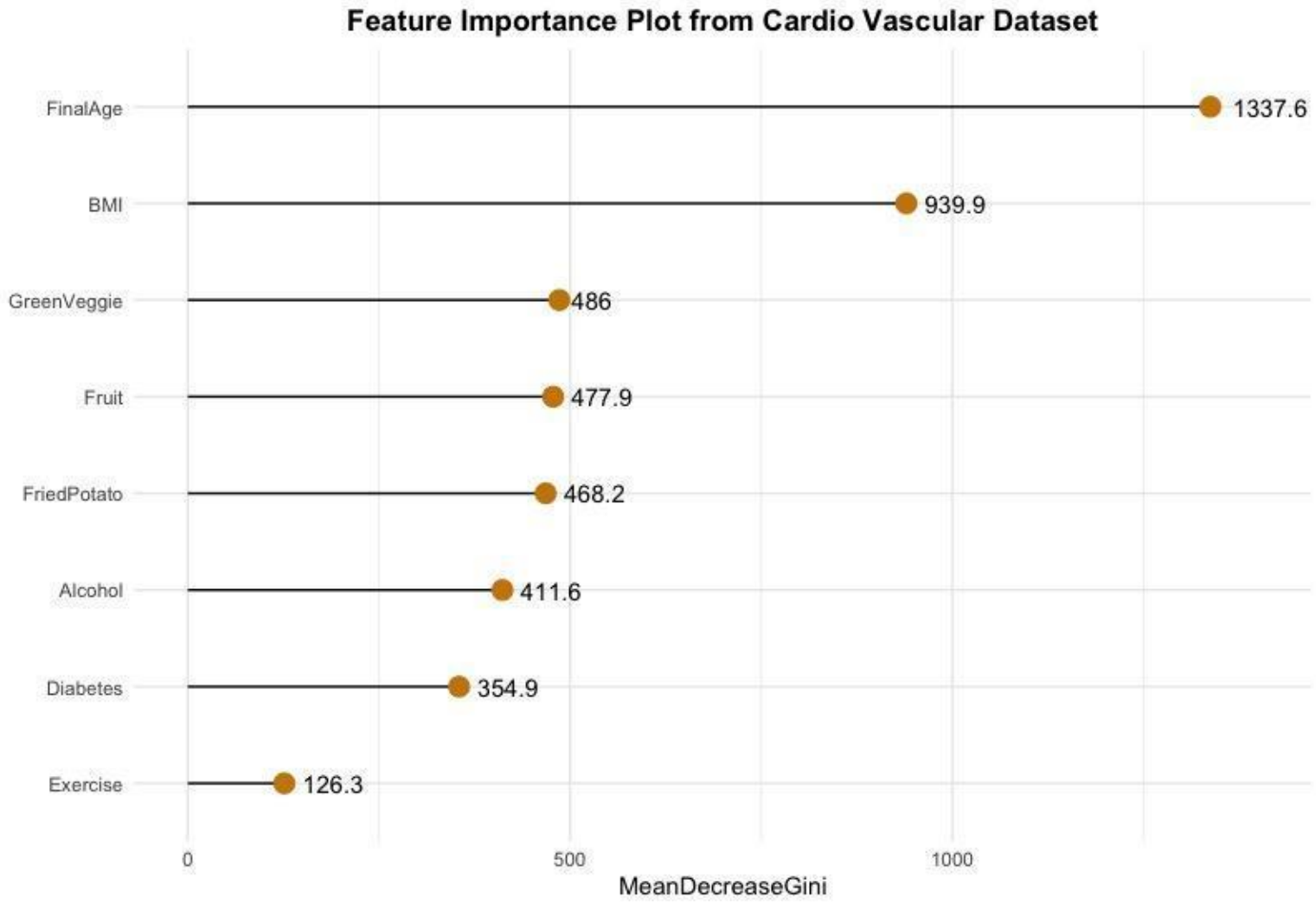
Model with manually selected Features

Feature Importance Plot

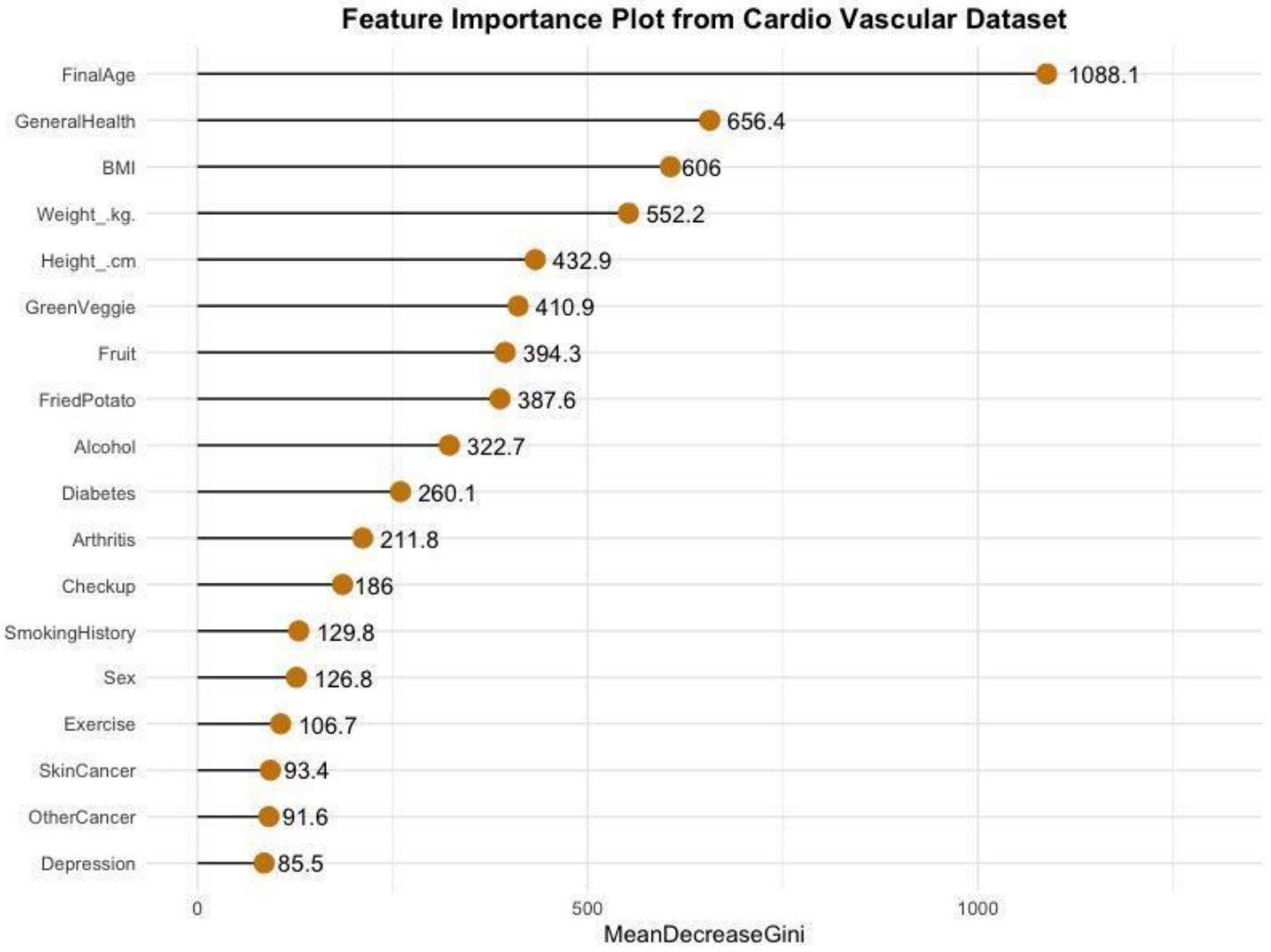
Interpretation:

- MeanDecreaseGini is a measure of predictive power of every feature based on Gini impurity criteria. Gini decrease was calculated for each variable across all trees in the forest to find its importance.

- Here, we can see this figure gives us a visualization of how useful each variable is in predicting heart disease. Interestingly, age and BMI were the most predicting variables followed by fruit consumption.

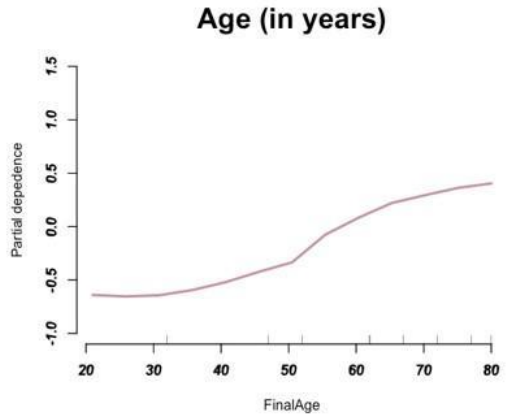


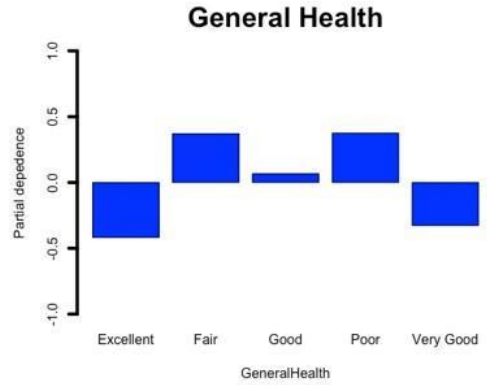
Model with All_Features

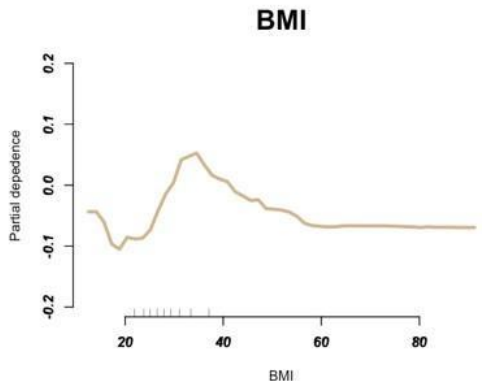


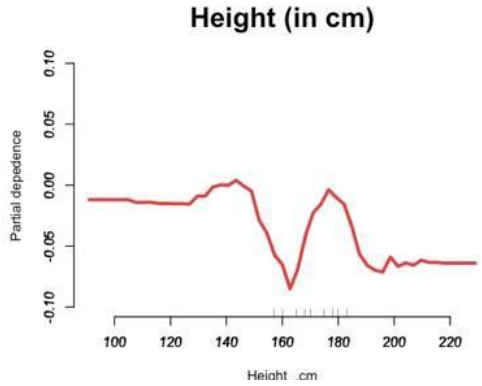
- Height and general health are importance features within the individuals with heart diseases.
- If we had not perfumed all variable analysis, we wouldn't have been able to observe that height could be an important predictor for heart disease.

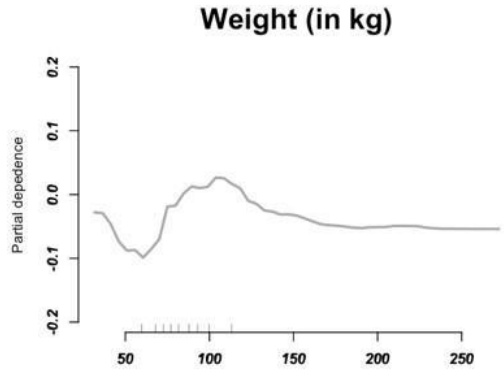
Partial Dependence Plot

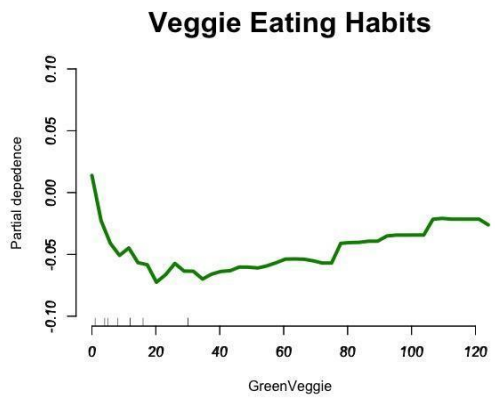


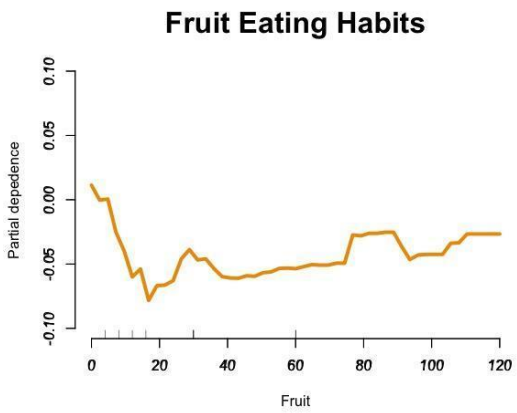


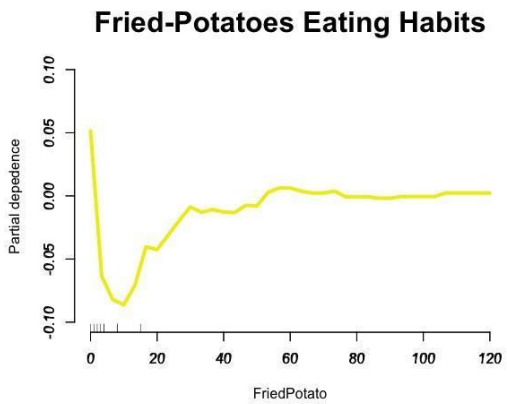












Discussion

Discussion—summary of the findings

Question 1.

- Important predictors were age, general health, BMI, Weight, Height, and Dietary habits like fruits, and vegetables.
    - In addition to BMI, the model found height and weight separately were also important predictors even though they are related to BMI.

Question 2

- By using random forest algorithm, we were able to analyze several key lifestyle factors that are important in predicting heart disease.
    - Model 1 (handpicked features) achieved 71% accuracy
    - Model 2 (with all features) achieved 74% accuracy (slightly better than Model 1).

Missed possibly important features such as General Health, and Height if we had not used all features model

However, how should we interpret height?

Implication

- Our finding shows the well-established links between lifestyle factors and cardiovascular health. Replicating these findings using other datasets and analytical approaches adds future credibility.
- It is also helpful in performing clinical practice.
- Our study shows how machine learning techniques can be used in addressing public health-related questions.

Limitations

- Although random forest is known for its robustness and ability to handle complex relationships, it may introduce the presence of irrelevant or redundant features that can dilute the signals from most informative variables hindering the ability of the model to identify the key factor of heart disease risk accurately.
- Outliers can have a substantial impact on the model.
- Handpicked features may be guided by practical experience or existing literature.

Future Direction

BMI, weight, fruit consumption, vegetable consumption included a few outliers in the positive direction.

- For future studies, we need to calculate and visualize the feature importance rankings for the new models without outliers. Compare these rankings with the original rankings from the models fitted on the dataset with outliers. Analyze whether the relative importance of features has changed significantly after removing outliers.
- By conducting this comparative analysis, we can gain insight into the robustness of our findings and the extent to which outliers may have influenced the interpretation of feature importance and partial dependence relationships.

Future Direction

Feature Engineering & Selection Refinement

- Exploration into domain-specific features related to cardiovascular health
  - genetic markers, advanced imaging data, additional lifestyle factors

Addressing Class Imbalance & Bias

- Explore techniques specifically to hand imbalanced datasets
  - oversampling, under sampling, etc.
- 20,000/300,000: perform analysis on whole dataset

Reference

- Lupague R.M.J.M., Mabborang R.C., Bansil A.G., Lupague M.M. (2023) Integrated Machine Learning Model for Comprehensive Heart Disease Risk Assessment Based on Multi-Dimensional Health Factors, European Journal of Computer Science and Information Technology, 11 (3), 44-58
- O'Rourke, R. A., Chatterjee, K., & Wei, J. Y. (1987). Cardiovascular disease in the elderly. Coronary heart disease. *Journal of the American College of Cardiology*, *10*(2 Suppl A), 52A–56A. https://doi.org/10.1016/s0735-1097(87)80449-7
- Spencer, E. A., Pirie, K. L., Stevens, R. J., Beral, V., Brown, A., Liu, B., Green, J., Reeves, G. K., & Million Women Study Collaborators (2008). Diabetes and modifiable risk factors for cardiovascular disease: the prospective Million Women Study. *European journal of epidemiology*, *23*(12), 793–799. https://doi.org/10.1007/s10654-008-9298-3
- Schramm, T. K., Gislason, G. H., Køber, L., Rasmussen, S., Rasmussen, J. N., Abildstrøm, S. Z., Hansen, M. L., Folke, F., Buch, P., Madsen, M., Vaag, A., & Torp-Pedersen, C. (2008). Diabetes patients requiring glucose-lowering therapy and nondiabetics with a prior myocardial infarction carry the same cardiovascular risk: a population study of 3.3 million people. *Circulation*, *117*(15), 1945–1954. https://doi.org/10.1161/CIRCULATIONAHA.107.720847
- Paffenbarger, R. S., Jr, Hyde, R. T., Wing, A. L., & Hsieh, C. C. (1986). Physical activity, all-cause mortality, and longevity of college alumni. *The New England journal of medicine*, *314*(10), 605–613. https://doi.org/10.1056/NEJM198603063141003
- Vella, C. A., Allison, M. A., Cushman, M., Jenny, N. S., Miles, M. P., Larsen, B., Lakoski, S. G., Michos, E. D., & Blaha, M. J. (2017). Physical Activity and Adiposity-related Inflammation: The MESA. *Medicine and science in sports and exercise*, *49*(5), 915–921. https://doi.org/10.1249/MSS.0000000000001179
- Sofi, F., Cesari, F., Abbate, R., Gensini, G. F., & Casini, A. (2008). Adherence to Mediterranean diet and health status: meta-analysis. BMJ, 337, a1344.
- Poirier, P., Giles, T. D., Bray, G. A., Hong, Y., Stern, J. S., Pi-Sunyer, F. X., ... & Eckel, R. H. (2006). Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss. Arteriosclerosis, thrombosis, and vascular biology, 26(5), 968-976.