Using random forest machine learning algorithms to identify a set of key lifestyle factors that uniquely distinguish individuals with and without heart diseases

Susmi Sharma, Christina Lam, Pratibha Pokharel

May 3, 2024

Class: EPPS6323 (Knowledge Mining)

The University of Texas at Dallas

Contributions:

Christina Lam: Introduction (*Literature Review* + *Background*), Discussion

Susmi Sharma: Introduction (*Kaggle Datasets* + *Random Forest Algorithm* + *Study Aims*), Method, Result, Discussion (*Summary of the Findings*)

Pratibha Pokharel: Introduction (*Background*), Discussion

Purpose

The primary objective of our study is to use machine learning algorithms to identify a set of key lifestyle factors that uniquely distinguish individuals with and without heart diseases.

Introduction

A. Literature Review

The importance of our project lies in its potential to revolutionize the prediction and prevention of cardiovascular diseases (CDV) by leveraging insights gained from analyzing diverse datasets, such as the Cardiovascular dataset (CDV) we have examined. By delving into the intricate relationship between various phenotypic traits and the presence of heart disease, we aim to uncover early indicators that could serve as valuable predictors.

Understanding the nuances of individuals' food habits, body mass index (BMI), and other lifestyle factors can offer crucial insights into their susceptibility to CDV. Research indicates that dietary patterns play a significant role in the development and progression of cardiovascular diseases. For instance, high intake of processed foods, saturated fats, and sugars is associated with an increased risk of heart disease, while diets rich in fruits, vegetables, and whole grains are linked to a lower risk (Sofi et al., 2008). Analyzing the dietary habits recorded in the CDV dataset can provide valuable information about the dietary patterns prevalent among individuals with heart disease, facilitating targeted interventions and dietary recommendations.

Furthermore, BMI, a measure of body fat based on height and weight, is a known risk factor for cardiovascular diseases. Research suggests that individuals with elevated BMI levels are more prone to developing heart disease, hypertension, and other related conditions (Poirier et al., 2006). By examining the BMI distribution within the CDV dataset and its correlation with

the presence of heart disease, we can identify high-risk populations and develop personalized strategies for weight management and cardiovascular risk reduction.

Moreover, the CDV dataset encompasses a wide range of features related to mental health, such as depressive episodes. Studies have demonstrated a bidirectional relationship between depression and cardiovascular diseases, with depression increasing the risk of developing heart disease and vice versa (Van der Kooy et al., 2007). By analyzing the prevalence of depressive episodes among individuals with heart disease within the CDV dataset, we can elucidate the interplay between mental health and cardiovascular health, paving the way for integrated interventions targeting both domains. So, this study is significant as it looks at all the different things that can cause heart problems and why we need to do something about it, focusing on the need for effective preventive measures and interventions.

In conclusion, our project holds significant importance for predicting and preventing cardiovascular diseases by harnessing the wealth of information contained within datasets like the CDV dataset. By uncovering associations between various phenotypic traits and the presence of heart disease, we can develop more accurate predictive models and tailored interventions to mitigate cardiovascular risk effectively.

B. Background

Heart disease encompasses a range of conditions affecting the heart, with coronary artery disease (CAD) being the most prevalent type in the United States. CAD impedes blood flow to the heart, increasing the risk of heart attacks which is a significant contributor to mortality according to the Centers for Disease Control and Prevention (CDC). Globally, cardiovascular disease (CVD) is the leading cause of death, taking approximately 17.9 million lives annually as reported by the World Health Organization (WHO). It remains a significant global health

concern, with its prevalence steadily increasing due to various lifestyle factors and behaviors. Familial predispositions significantly influence heart disease susceptibility, with individuals possessing a family history of heart disease at heightened risk. Various forms of heart disease and related conditions, including high blood pressure and high blood cholesterol, can exhibit familial clustering, reiterating the importance of understanding one's family health history in preventing future cardiac events, as emphasized by the CDC.

While traditional risk factors such as hypertension, diabetes, obesity, and smoking have long served as predictors for heart disease, recent research has unveiled a concerning trend. Studies conducted by the University of Sydney and Heart Research Australia reveal an increase in heart attack patients lacking conventional risk factors, such as elevated cholesterol levels. This observation highlights the complexity of heart disease causes and the need for a more comprehensive understanding of contributing factors. Despite advancements in cardiovascular research, a unified theory explains that the causes of heart disease is difficult to define.

Accurately characterizing the important lifestyle factors of patients with and without heart diseases is pivotal. This will enable clinicians to provide precautions to people who are at risk. Early detection and management of CVD risk factors are crucial for preventing disease progression and reducing the burden of cardiovascular events. Furthermore, findings of such studies may allow policy makers to come up with policies that discourses high risk factors associated with heart diseases. A possible approach to addressing questions around phenotypically complexity of cardiovascular health is through the use of machine learning tools that learn statistically relevant features within this complexity. Such machine learning methods recently have emerged as a well-suited technique to explore the risk factors associated with heart disease.

In recent years, significant strides have been made in cardiovascular medicine, exemplified by breakthroughs in technology aimed at restoring blood flow to obstructed and narrowed arteries. These advancements, highlighted by the American Heart Association, hold promising potential in preventing deaths and disability across diverse patient populations, including those with severe comorbidities. As researchers deepen their understanding of the complex nature of cardiovascular disease, the search for effective interventions to combat heart disease and stroke persists, driving progress towards improved patient outcomes and enhanced public health initiatives. By integrating diverse approaches into comprehensive cardiovascular care models, healthcare providers can better address the complex and multifactorial nature of cardiovascular disease, ultimately leading to improved patient outcomes and population-wide cardiovascular health.

C. Kaggle CVD Datasets

A variety of heart-health related survey data capturing lifestyle factors of individuals is publicly available, such as on Kaggle (refer to the CVD Dataset Link). The Cardiovascular Disease Risk Prediction Dataset (CVDs) represents one of many such datasets and includes key variables potentially indicative of cardiovascular diseases. This dataset originates from a telephone-based survey conducted in 2021 across 47 states, using the Behavioral Risk Factor Surveillance System (BRFSS) administered by the Centers for Disease Control and Prevention. For further details on the questionnaire, interview procedures, and the data itself, please see the BRFSS 2021 Overview. While the comprehensive BRFSS dataset encompassed 304 features from 438,693 participants, the version available on Kaggle has been condensed to include only 19 features that provide insights into the mental and physical health, dietary habits, and lifestyle choices of the patients.

For our class project, we utilized the CVD dataset to investigate whether we could effectively classify participants with heart disease from those without, leveraging the diverse lifestyle factors provided in the data. Our primary goal is to develop a model with high predictability. Additionally, we seek to identify the key features that significantly contribute to distinguishing individuals with and without heart disease. By unraveling the complex relationship between various phenotypic traits associated with heart diseases, we aim to gain insights that could potentially revolutionize the prediction and prevention of cardiovascular diseases. Ultimately, this knowledge may empower researchers and clinicians to proactively steer their patients away from the risk of heart disease.

To explore the above question using the CVD dataset, we considered several machine learning algorithms. Previous analyses of this dataset have utilized various statistical and machine learning techniques to uncover patterns and relationships within the data. These methodologies include logistic regression modeling, which aims to predict binary outcomes based on predictor variables as well as univariate, multivariate, and bivariate analyses, which explore relationships between individual and multiple variables. Additionally, deep learning approaches such as deep neural network (DNN) models, recurrent neural network (RNN) models, and long short-term memory (LSTM) models have been used to capture complex nonlinear dependencies and temporal dynamics within the data. The CVD data features include twelve continuous features and seven categorical, with several binary predictors such as smoking status, diabetes presence, and depressive episodes, which are relevant for heart health assessment. To us, 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 for predicting heart disease among subjects in this dataset. However, delving deeper into Kaggle

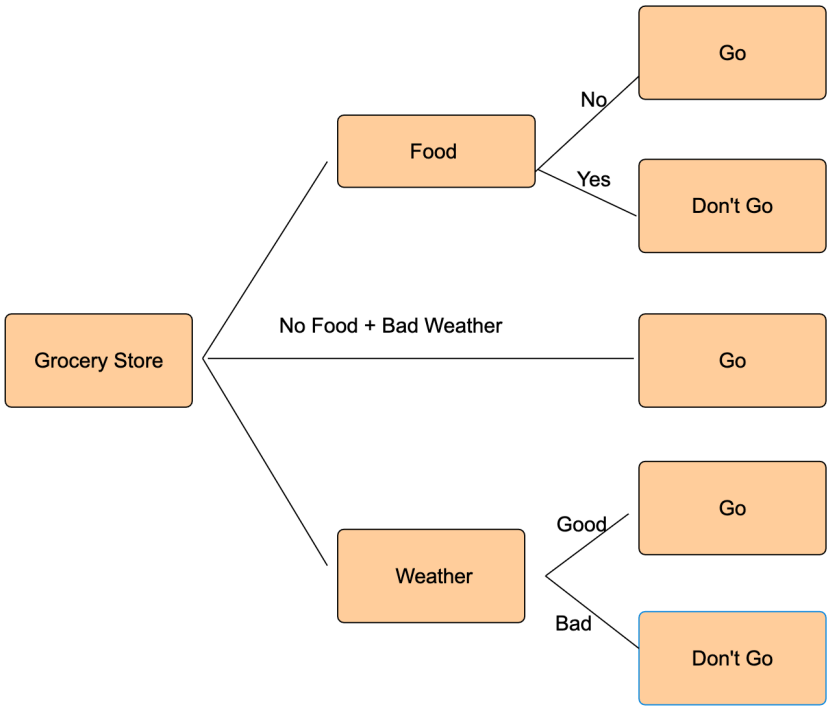
website and prior research articles on the dataset, we found out that logistic regression was previously utilized on the dataset (Lupague et al, 2023; Kaggle).

Consequently, random forest was the machine learning algorithm that seemed suitable for our project. Although we found one Kaggle article on the CVD dataset incorporated random forest algorithm, they only compute the performance and predictive power of the random forest model (see Random Forest Kaggle). They did not explore the key characteristics of patients with heart diseases using the random forest model. Furthermore, this and other analysis on the dataset in Kaggle was done using the python programming language. We utilized novel random forest modeling techniques to its fuller capacity on the CVD dataset in R. In this regard, our ideas, codes and approach all stand out for its originality.

D. Machine Learning Algorithm: Random Forest

The concept of random forest modeling is an advanced extension of decision tree modeling that incorporates multiple decision trees to aggregate their predictions, thereby enhancing predictive power and stability. This method mirrors systematic human thought processes, where decisions are made based on a series of sequential steps. For example, consider the decision to visit a grocery store, influenced by factors such as weather, personal schedule, and the current amount of food in the refrigerator. In decision tree modeling, this process would be emulated algorithmically, with data features evaluated at each step to predict outcomes, such as deciding to visit the store if the weather is pleasant, or refraining if there's enough food. The following figure (decision tree 1) shows how we would make decisions on whether to go to grocery stores or not.

Decision Tree 1.



Unlike a single decision tree, which follows one analytical path, a random forest utilizes numerous trees, each analyzing a variable set of decision pathways. This incorporation of inherent randomness and additional degrees of freedom helps to mitigate overfitting, a common pitfall associated with single decision trees. The ensemble nature of random forests not only improves the generalizability of the model but also allows for a more thorough examination of the significance of each feature in classifying outcomes like heart disease. This approach effectively uses collective insights to provide a more in depth understanding of complex phenomena.

A distinctive feature of the random forest algorithm is its ability to evaluate the importance of each variable. In this context, variable importance indicates how significantly a variable has contributed to diminishing the error rate or enhancing the model's predictive accuracy. This assessment facilitates the ranking of predictors from most to least crucial, aiding in more informed decision-making and prediction improvements. Additionally, partial

dependence plots can be utilized to examine the influence of each feature on the model, across various levels of the feature values. Therefore, in order to improve our understanding and interpretability of this model, finally, we also plotted partial dependence plots on the top 8 important features.

E. Aim of the Study

To encapsulate once again, our study attempts to characterize the identifying lifestyle aspects of individuals with heart diseases using random forest machine learning algorithms. Overall, our approach should offer insights on how the data and machine learning algorithms should be handled or incorporated while investigating health-care related research questions.

Method

A. Data

The CVD data did not include any missing data. About 8% (n = 24, 971) of the total participants in the telephone survey had reported having heart diseases.

B. Data Pre-processing

Despite having access to a large dataset of observations, we opted to analyze only a subset of the CVD data. This decision was driven by the considerable time random forest required to construct a model when using the entire dataset. The algorithm's reliance on numerous decision trees demands substantial computational resources. Consequently, we mitigated this issue by reducing the dataset to 20,000 randomly sampled data points. This subset comprised an equal distribution of observations from both groups: 10,000 participants with healthy hearts and 10,000 with unhealthy hearts.

During the next stage of pre-processing, we refined the variables age and diabetes for modeling purposes. Since each participant's age in the CVD dataset was represented as a range rather than a numerical value (e.g., "50-60" and "80+"), we converted these age bins into numerical values by assigning each participant's age to the midpoint of the range they fell into. The variable "diabetes" encompassed four distinct levels of observations: "Yes," "No," "Yes, but females reported only during pregnancy," and "No, pre-diabetes or borderline diabetes." Notably, there were a small number of observations in the last two conditions. To maintain clarity and prevent confounding interpretations during analysis, we excluded 677 observations that fell outside of the "yes" and "no" categories for diabetes. As a result, our model utilized 19,323 observations, comprising 9,649 participants with heart diseases and 9,674 participants without heart diseases.

Finally, the CVD dataset included many of the dichotomous variables (such as exercise, heart disease, and depression) in the character form. We then converted the datatype of several of these variables into factor form, so R reads them (yes/no) as categorical entities, rather than as characters.

C. Features for the Model

We included all the 19 features provided in the dataset, letting the model extract and use whichever features it finds relevant in predicting the class of the participants with or without heart disease. We calculated the models' performance for both models and observed their predictive power as well as out-of-bag error estimates. In addition to finding the important measures of heart disease, our comparative study attempts to provide pros and cons of handpicked theory-driven vs. machine learning data-driven techniques. As the Cardiovascular Disease Risk Prediction Dataset clearly provides categorical labeling (0 and 1) for individuals

with and without heart disease, we use those labels as the output of our tree-based supervised learning algorithms.

Table1.

Random Forest Model			
Model	**Features**	**Input**	**Output**
1	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)

D. Random Forest Model Algorithm

The random forest algorithm begins through the selection of data subsets through randomized replacement sampling techniques of the entire dataset that ensures representative sample distributions. It then constructs a decision tree for each of these subsets, generating a "forest" of decision pathways that associate phenotypic observations with presence, absence, or degree, of a set of features. The process of replacement sampling or bagging attempts to simulate the way data may occur in real populations, where some features and associated phenotypes are more likely to be observed in natural settings than others. The combinations of features and phenotype data that are not sampled as often are referred to as "out of bag data" (OOB). From here, the algorithm proceeds to construct a decision tree for each subset of training data. A notable aspect of random forest modeling is its inherent feature selection that comes about through the random selection of a feature or random features at each node of the tree. Each tree's

growth continues to propagate until the process can no longer split the tree. Once all possible trees that can be made from the data are completed through bootstrapping the dataset, predictions for out of the bag data is formulated by aggregating (majority vote) the decisions across all trees. This methodology ultimately enhances the random forest model's resistances to overfitting and improves its generalizability.

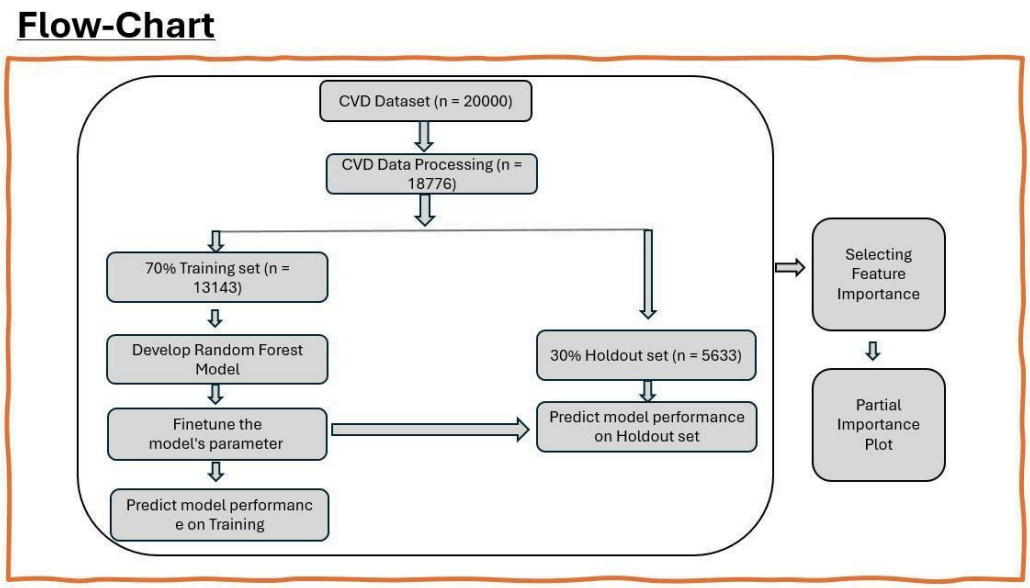
Steps to Random Forest Model Training

1. Data Splitting: We begin our classification process with a systematic partitioning of the data from both the heart-healthy and the individuals with heart disease into two subsets: a larger training set, which accounts for 70% of our data, and a smaller testing set, making up the remaining 30%.
2. Model Developing: We then fit the suitable model to the training data, developing a model that discerns the probability of heart disease for each participant. At this stage, we also recorded the model's out of bag error estimate and its predictive power.
3. Tuning Parameter: Random forest models require a set of crucial parameters tuning. We optimized two of them—first, the total number of decision trees (ntrees), and second, the number of features to consider splitting at each node of the decision trees(m). While the default number of trees that the random forest package sets to is 500, we investigated whether increasing the number of trees to 1000 would lead to a noticeable improvement in the prediction rate. Similarly, to determine the optimal number of variables for splitting, we further examined how the error rate varies as we increase the number of features splitting.
4. After the parameter tuning procedure, the model is finally tasked with predicting outcomes for the individuals in our testing set. The accuracy of the model is evaluated by

how well these predicted probabilities match up with the participants' actual diagnostic outcomes.

5. Finally, the important features used in the classification were noted for the model and were analyzed briefly voa partial dependence plots.

The above steps are also illustrated in the flow-chart below.



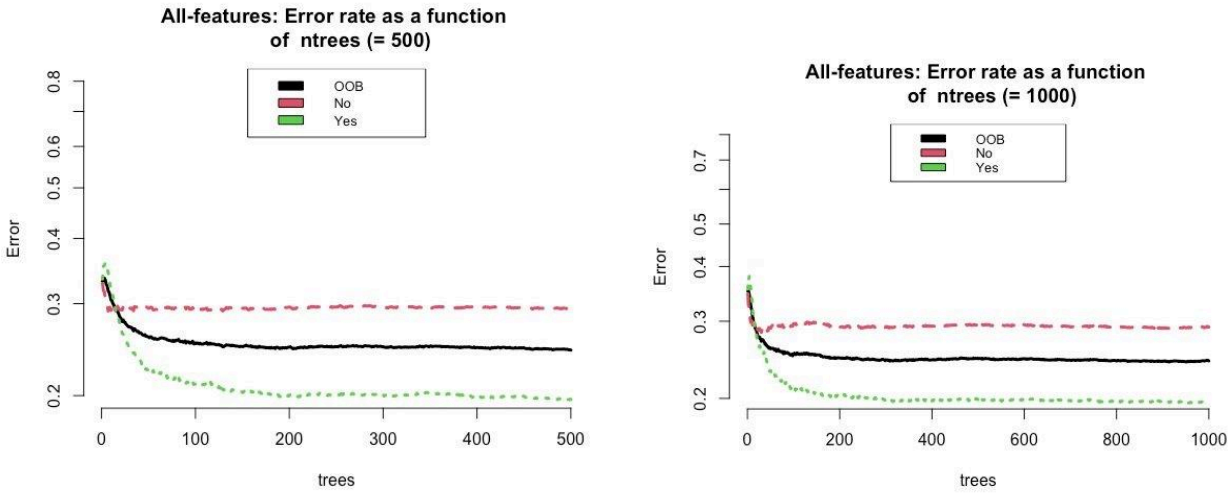
E. Computational Programming Software

For our statistical computations, we will use R as the primary software due to its robust packaging ecosystem and capacity for handling complex statistical tasks. “randomforest”, “caret”, “ggplot2”, and “ggalt” were some of the packages extensively used in our project.

Results

A. Tuning Parameter

As discussed in the method section, we first utilized the random forest algorithm to the training data set and obtained our model. Then, before analyzing the model, we ensured that two important tuning parameters (ntrees)—number of decision trees in the forest and number of splits considered by each tree while splitting the node (m)—were set to their optimal values, since the default number of decision trees, which is 500, may not be the optimal number of decision trees. Figure 1 shows how the model's out of the bag error estimate changed as a function of number of decision trees in the forest when ntrees is 500 vs 1000.



As seen in the figure, increasing the number of decision trees from 500 (figures on left) to 1000 (figures on the right) did not substantially decrease the out of bag error (OOB) estimate. Consequently, we fixed our number of decision trees for our model to their default value of 500, as increasing the complexity of the model did not appear to improve the model performance. 500 decision trees were enough to get a stable out of bag error estimate (OOB).

Then, the performance of the model was analyzed at different variable split numbers (m) from 1 to 10. While the default value of m used in randomforest package is $\sqrt{p}$, where p is the number of predictors, our analysis showed that the variable split number of 2 was capable of getting the lowest OOB for our model.

B. Model Performance

After parameter tuning, the performances were computed for the model. Table 2 illustrates the confusion matrix performance on both the training and testing set. The out of bag error estimate for the model was 24.43%.

Table 2: Model with All Features

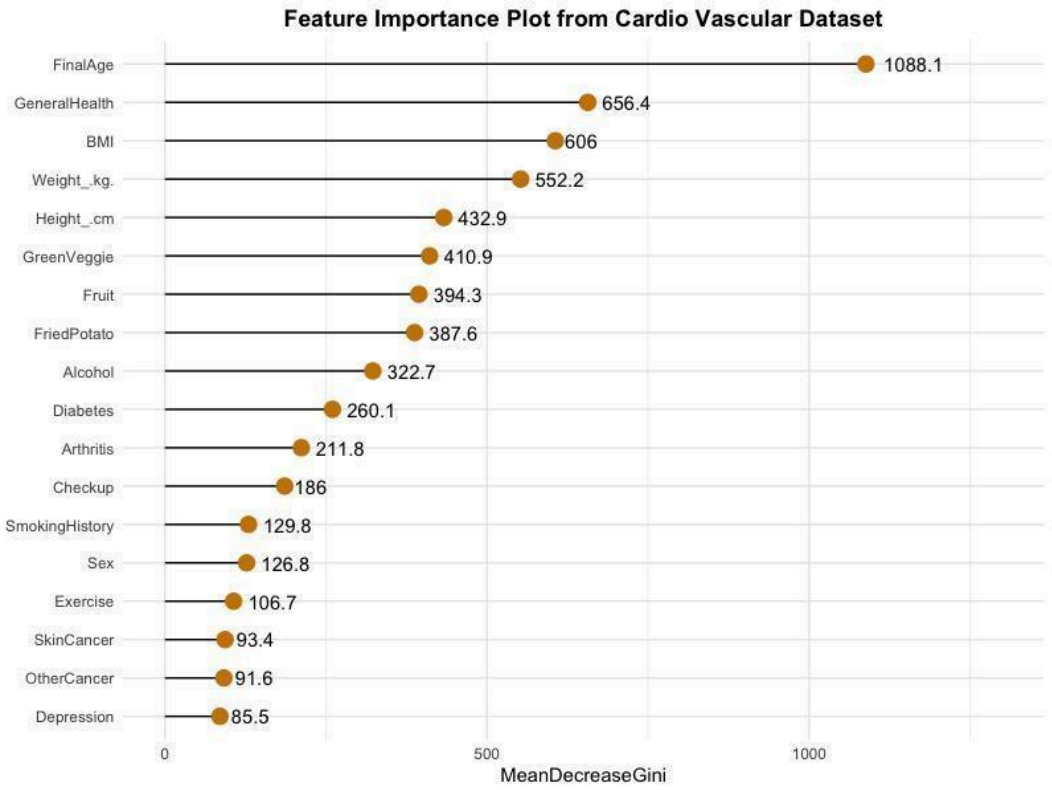
Predicted Observation	**True Observation**				
	Training Set			Testing Set	
		Heart Healthy	Heart Disease	Heart Healthy	Heart Disease
	Heart Healthy	6091	384	2022	541
	Heart Disease	657	6394	904	2330

The specificity and sensitivity of our random forest model were calculated to assess its accuracy on the testing dataset. Our result indicated that our model with all features achieved a specificity of 81.16% and sensitivity of 69.90% in detecting heart condition in adult participants. According to our analysis, the specificity of both models was higher compared to their sensitivity. This suggests that our model was more successful in correctly rejecting the participants without heart disease than in correctly identifying participants without heart disease.

C. Feature Importance

Finally, we computed the rankings of the feature importance, from the ones that contributed the most to the ones that contributed the least, in determining participants with and without heart disease. Figures 2 illustrate the contributions of each of the features to model performance for our model respectively according to mean decrease in gini measure. Gini captures how much each of the variables contributes to the similarity of the nodes and leaves in the random forest. Then the mean decrease in gini algorithm calculates how much of the similarity or homogeneity of the nodes in a tree is reduced as a function of the individual feature while making decisions. In other words, this algorithm gives us the order of the importance of the features in determining whether participants have heart disease or not.

Figure 2



Our random forest models showed that patients' age is the most important factor in distinguishing people with and without heart disease. According to our model, after age, one's general health is the next important measure to predict their heart condition. Then, it was followed by BMI, weight, height. and their eating habits. Smoking history, sex, exercise, skin cancer, other cancer and depression were the least important predictors in the model.

D. Partial Dependence Plot

To improve the interpretability of the study, in the next step of the analysis, we plotted the partial dependence plot of the top 8 features in Figure 3. This allowed us to examine how exactly these features contributed to the prediction of patients' heart conditions. The following partial dependence plots show how the probability of having heart disease changes at the different values/levels of the predictors. For instance, our random forest model predicted that the individuals are more likely to get heart disease as they get older, though the probability does not increase exactly linearly.

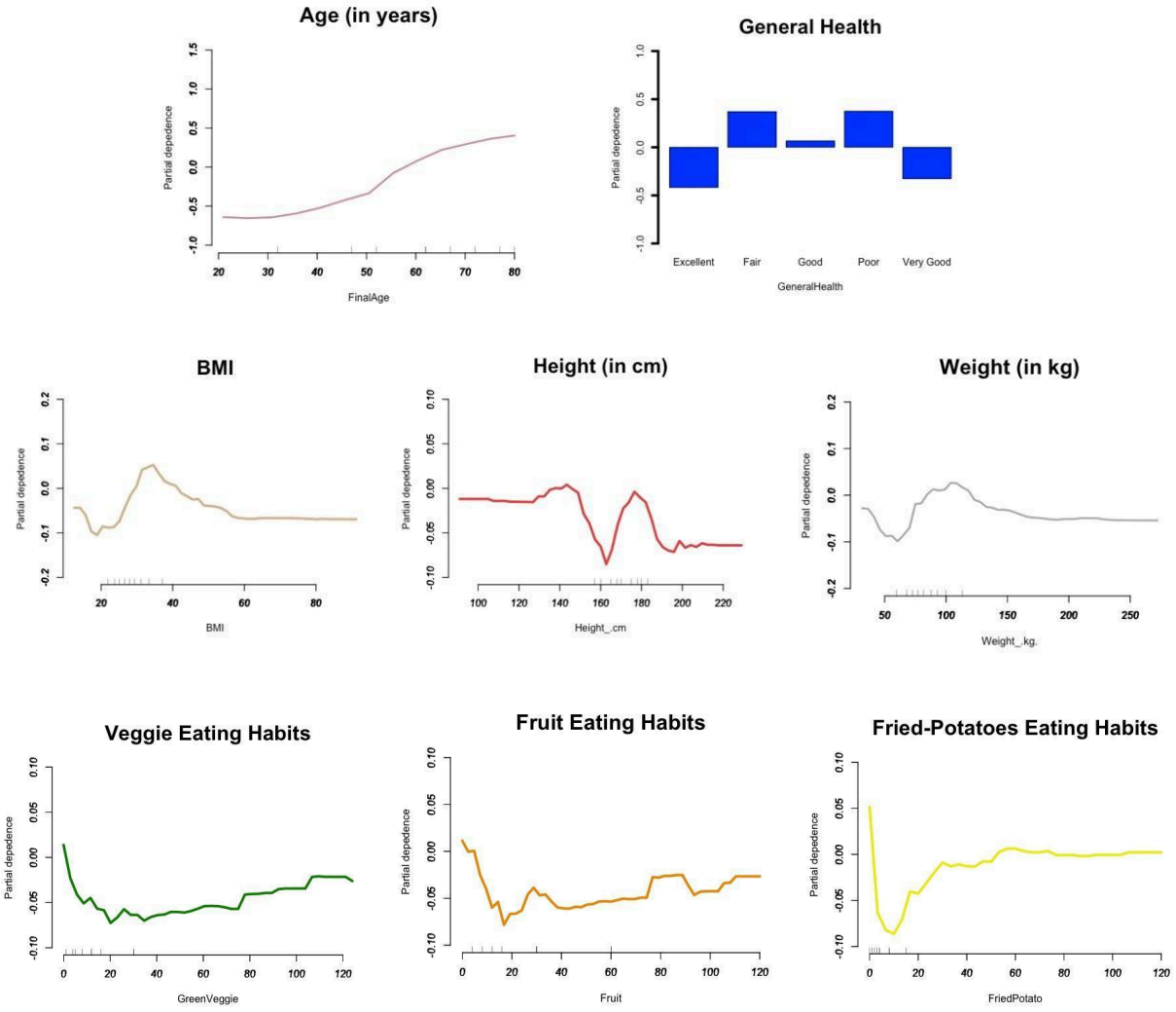


Figure 3: Partial dependence plot of the top 8 features from the model. The y axis represents the patient's predicted rate of heart-disease, whereas the x axis represents the value of the predictor. For first plot (1), it is age in years, levels on general health (2), BMI index (3), height in cm (4), weight (in kg) (5), servings of veggie consumption servings in a month (6), serving of fruits in a month (7), and fried potatoes consumption in a month (8). (*Note: some of the features include values outside the decile range*).

The plot suggests that predicted heart rate goes up with the increase in age. However, interpreting other partial dependence plots requires caution. Looking closely at the partial dependence plot, we observed that height, weight, and dietary predictors are not normally distributed and they may include a significant proportion of outliers. However, if we only observe the data where decile points lie proportionally, the plots show that the rate of heart

disease increases with BMI, height, and weight, but decreases when people have good general health and they consume fruits, vegetables and even fried potatoes. Interpreting how exactly fried potatoes affect the rate of heart diseases however requires further analysis.

Discussion

A. Summary of the Findings

As shown in the result section, random forest Model reached 74% accuracy in accurately predicting the heart condition of patients on an independent dataset. Examining the feature importance ranking of the full feature model, we discovered that patients' age, their general health, BMI, weight, height, and dietary habits were the most significant and informative predictors, whereas smoking habits, sex, exercise, skin cancer, other cancer and depression were not very informative in predicting one's heart condition.

According to our model, after around 56 years of age, the likelihood of patients getting heart diseases increases almost linearly. Next, people with overall good health were predicted with less risk of heart disease. Our model also validated numerous previous findings on the positive relationship between BMI and heart disease. For instance, excess body weight, especially when concentrated around the abdomen, has been linked to various cardiovascular risk factors such as hypertension, dyslipidemia, insulin resistance, and inflammation, all of which contribute to the development of heart disease (Angelantonio, 2016). Moreover, our model suggested an inverse relationship between fruit consumption and green vegetables consumption with the risk of heart disease, whereas a direct relationship between fried potato consumption and heart disease. These findings are consistent with current research findings, as research has shown that fruits, rich in antioxidants, vitamins, minerals, and fiber, have cardioprotective effects. Regular consumption of fruits is associated with lower blood pressure, improved lipid

profiles, reduced inflammation, and enhanced endothelial function, all of which contribute to a lower risk of heart disease (Wang, 2014; Boeing, 2012). On the other hand, high consumption of fried potatoes, especially in the form of French fries and potato chips, has been associated with an increased risk of heart disease. Fried potatoes are typically high in unhealthy fats, calories, and sodium, and their consumption is linked to weight gain, elevated blood pressure, dyslipidemia, insulin resistance, and inflammation, all of which are risk factors for heart disease (Mozaffarian, 2011).

Finally, constructing a random forest model using all features of the CVD data, we discovered that machine learning classification technique has an amazing capability of showing the relationship between predictors and the outcome that would have been missed otherwise. While it was reasonable that age, general health, BMI and dietary habits played a crucial role, we also found that in addition to BMI, height and weight were also incorporated as the important predictors in our model. Even though height does not seem to play a key role in explaining heart disease, our feature importance plot suggests that it *is* one of the highly predictive features. This suggests that there may be a direct/indirect correlation between increased height and risk of heart disease, prompting the need for further investigation in this area.

B. Implications of the Findings

The findings from our study can directly inform clinical practice by providing healthcare professionals with actionable information to counsel and educate their patients. By highlighting factors such as age, BMI, and dietary habits, clinicians can tailor their advice and recommendations to address these modifiable risk factors. We found that after around 56 to 80 years of age, the likelihood of patients getting heart diseases increases almost linearly. Maybe clinicians can use this information by asking patients after 50 years of age to consistently do

physical exams and their heart condition. We found a direct relationship between general health and heart health. Clinicians can use this information by recommending patients with poor general health to keep track of health.

Then, our model suggests that as weight (between 50 kg to 100 kg) and height (between 160 cm to 180 cm) increases, the rate of getting heart disease seems to increase proportionally. Though this finding needs further investigation before drawing a conclusion, there is a possibility of a genuine relationship between these factors . If a true relationship exists, this is a significant finding, since policy makers can target and designate resources and special heart-health plans differently for people from different countries and ethnicities, since distribution of heights for people from different countries and different ethnicities vary. Moreover, health care workers can also provide specific guidance on incorporating more fruits and vegetables into patients' diets while limiting the consumption of fried or processed foods. Overall, our study underscores the need for regular monitoring and tracking of these lifestyle factors.

Our findings also align with and reinforce the existing knowledge about the risk factors associated with heart disease. The result corroborates the well-established links between lifestyle factors and cardiovascular health. Replicating these findings using a different dataset and analytical approach (random forest), we were able to strengthen the hypothesized relationship between heart disease and life-style factors. Moreover, by pointing out age, general health, BMI and dietary habits play a key role in determining heart condition, we demystified the puzzle to heart disease. Lifestyle factors play a significant role and they should be incorporated in order to protect ourselves and our loved one from heart diseases. Furthermore, the ranking of feature importance provides valuable insights into the comparative contributions of each factor, which can inform targeted interventions and preventive strategies.

Finally, the successful application of the random forest algorithm in our study showcases its potential for exploring other complex public health issues. Similar techniques can be employed to analyze diverse datasets, identify patterns, and uncover critical risk factors or predictors associated with various health conditions or outcomes. Our comparative approach highlights the versatility of machine learning techniques. Demonstrating the effective machine learning methods used in our study paves the way for further adoption and exploration of these techniques in public health research.

C. Future Direction

For future studies, we will ask if the predictive performance of the model increases if we remove the outliers from all of our features. It is possible that the 74% accuracy of the model was due to the inclusion of the outliers in the variables. Then, we will calculate and visualize the feature importance rankings for the new models without outliers. We will compare these rankings with the original rankings from the models fitted on the dataset with outliers and 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. If the feature importance rankings and partial dependence plots remain consistent after removing outliers, it would suggest that the original findings are robust and not overly influenced by extreme values.

Moving forward, then, we will also explore the relationship between height, weight and heart disease. Is there a genuine relationship between height and heart disease? Or the observed association between height and heart disease is due to the interaction between height and weight? In other words, we will analyze if the effect of height on heart disease changes across different

levels of weight. This analysis will help us identify whether height has a direct and unique relationship with heart disease.

Because the dataset was so large with over 300,000 features, only 10,000 heart healthy and 10,000 heart unhealthy features were analyzed. So in future studies we will perform the same analysis on the whole dataset and observe if we replicate our findings. Class imbalance within the CDV dataset presents a significant challenge that warrants attention in future endeavors. To mitigate the potential biases introduced by uneven class distributions, we may explore techniques such as oversampling, undersampling, or the application of advanced ensemble methods specifically made to handle imbalanced datasets. By ensuring equitable representation of both positive and negative instances of heart disease and using the entire dataset, there may be improvements in the accuracy of the model as well as in the robustness and generalizability of the models across diverse patient populations.

Finally, we will analyze whether selecting a subset of the feature or a different set of features enhances the accuracy and interpretability of the models. While random forests offer inherent feature selection capabilities, further exploration into domain-specific features related to cardiovascular health, such as genetic markers, advanced imaging data, or additional lifestyle factors, could enrich the predictive models. Additionally, techniques such as principal component analysis (PCA) or recursive feature elimination (RFE) can be employed to identify the most informative features and mitigate the impact of irrelevant or redundant variables.

D. Limitation of our Study

While the study employed a random sampling approach to create a manageable subset of the CVD dataset, there is a possibility that the results may not fully generalize to the entire population. By analyzing only a portion of the data (20,000 observations), our findings may not

have captured the complete variability and nuances present in the larger dataset. It is also worth noting that a larger, but biased or non-representative sample may not necessarily yield more generalizable or interpretable results compared to a smaller, well-curated dataset that accurately reflects the population of interest.

Although random forests are known for their robustness and ability to handle complex relationships, they are not immune to certain limitations. The model also may introduce additional complexities such as the presence of irrelevant or redundant features that can dilute the signal from the most informative variables, potentially hindering the model's ability to identify the key determinants of heart disease risk accurately.

Outliers, which are observations that deviate significantly from most of the data, can have a substantial impact on the performance and interpretability of machine learning models, including random forests. Many features of our dataset such as height, weight, etc may contain influential outliers, which may have skewed the feature importance rankings and affected the partial dependence plots, potentially leading to biased or misleading interpretations.

Reference

Lupague, R.M.J.M., R.C. Mabborang, A.G. Bansil, and M.M. Lupague. 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., K. Chatterjee, and J. Y. Wei. 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., K. L. Pirie, R. J. Stevens, V. Beral, A. Brown, B. Liu, J. Green, G. K. Reeves, and 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., G. H. Gislason, L. Køber, S. Rasmussen, J. N. Rasmussen, S. Z. Abildstrøm, M. L. Hansen, F. Folke, P. Buch, M. Madsen, A. Vaag, and C. Torp-Pedersen. 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, R. T. Hyde, A. L. Wing, and C. C. Hsieh. 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., M. A. Allison, M. Cushman, N. S. Jenny, M. P. Miles, B. Larsen, S. G. Lakoski, E. D. Michos, and M. J. Blaha. 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., F. Cesari, R. Abbate, G. F. Gensini, and A. Casini. 2008. "Adherence to Mediterranean Diet and Health Status: Meta-analysis." *BMJ* 337: a1344.

Poirier, P., T. D. Giles, G. A. Bray, Y. Hong, J. S. Stern, F. X. Pi-Sunyer, et al. 2006. "Obesity and Cardiovascular Disease: Pathophysiology, Evaluation, and Effect of Weight Loss." *Arteriosclerosis, Thrombosis, and Vascular Biology* 26 (5): 968-976.

Van der Kooy, K., H. van Hout, H. Marwijk, H. Marten, C. Stehouwer, and A. Beekman. 2007. "Depression and the Risk for Cardiovascular Diseases: Systematic Review and Meta-analysis." *International Journal of Geriatric Psychiatry* 22 (7): 613-626.

Supplemental Links

https://www.cdc.gov/heartdisease/about.htm
https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
https://www.cdc.gov/genomics/disease/fh/history_heart_disease.htm
https://www.sydney.edu.au/news-opinion/news/2017/08/30/heart-attacks-with-no-obvious-risk-factors-on-the-rise.html
https://www.heart.org/en/around-the-aha/aha-names-top-advances-in-cardiovascular-disease-research-for-2023
https://www.cdc.gov/brfss/annual_data/2021/pdf/Overview_2021-508.pdf

https://www.cdc.gov/nutrition/downloads/Data-Users-Guide-BRFSS-Fruit-and-Vegetable-Questions-508.pdf

Also, the first decision tree template figure was drawn using this website.