Research of Influence Factors that Possibly Lead to Cardiovascular Disease using Machine Learning

Huailang Peng*

Department of Mathematics, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China

*Corresponding author: Huailang.Peng24@student.xjtlu.edu.cn

Abstract. While previous studies have studied and demonstrated that the incidence of cardiovascular disease (CVD) is related to various factors such as hypertension, high cholesterol, smoking, diabetes, obesity, lifestyle, pregnancy and so on, there still exist many unidentified factors that are valuable to be researched on. This research tries to apply three classical machine learning algorithms to deal with the data from the Kaggle website. The dataset was compiled by Alphiree from the online Cardiovascular Diseases Risk Prediction Dataset. This dataset cited data from 2021 Behavioral Risk Factor Surveillance System (BRFSS). This study uses and processes the 5,523 records collected as data from the BRFSS in 2021 from World Health Organization (WHO). It is concluded that the BMI, Weight, Age Category, Height, Green Vegetables Consumption, Fruit Consumption and FriedPotato Consumption have relatively strong relationships with the development of CVD, while General Health, Checkup, Exercise, Skin Cancer, Other Cancer, Depression, Diabetes, Arthritis, Sex, Smoking History and Alcohol Consumption have relatively weak relationships with having a CVD. This result provides some new perspectives to study the pathogenesis and treatment of CVD and point to the way for further research afterward.

Keywords: Cardiovascular disease; logistic regression; naive Bayes; random forest; influencing factors.

1. Introduction

Cardiovascular disease (CVD) is believed to be a leading reason of death both in the United States (US) and globally [1]. Hence the importance of studying the factors that affect CVD and assessing CVD risk should be valued. Fuster-Parra, Pilar, et al. posited that conducting an extensive, thorough research of cardiovascular risk factors (CVRF) appears to be of paramount importance in the research of cardiovascular disease, with the aim of preventing (or mitigating) the likelihood of developing or dying from CVD [2].

The factors leading CVD are really complicated. Dunn et al. argued that counseling on lifestyle physical activity is equally efficacious as structured exercise programs in enhancing physical activity levels and improving CVD risk variables among initially sedentary males and females, following a six-month duration [3]. Benschop et al. researched on Cardiovascular risk factors after pregnancy such as chronic hypertension, renal dysfunction, dyslipidemia, diabetes and subclinical atherosclerosis [4]. Although preceding studies have demonstrated that the morbidity of cardiovascular disease correlates with various factors such as hypertension, high cholesterol, smoking, diabetes, obesity, lifestyle, pregnancy and so on, there still have many unidentified factors that worth further studying. There are many factors related to CVD that were not taken into consideration. Therefore, this paper focuses on 18 variables, namely General Health, Checkup, Exercise, Skin Cancer, Other Cancer, Depression, Diabetes, Arthritis, Sex, Age Category, Height, Weight, BMI, Smoking History, Alcohol Consumption, Fruit Consumption, Green Vegetables Consumption, and Fried Potato Consumption, as well as their correlations and respective weights in relation to CVD.

In the similar vein regarding CVD. Peltola, Tomi, et al. advocated that enhancing disease risk prediction constitutes a primary objective in epidemiological research. Non-communicable diseases, a significant portion of which develop and progress gradually, are a major cause of morbidity globally. Precise risk prediction can be utilized to screen individuals for targeted interventions [5]. Chekouo et al. conducted rigorous Bayesian integrative analysis and predictive research, with specific application to atherosclerosis cardiovascular disease [6]. Elsayad et al. used Bayesian classifiers to diagnose

cardiovascular diseases [7]. Miranda and her colleagues employed the Naive Bayes classifier to ascertain the level of risk associated with cardiovascular disease for adults [8]. Ambrish et al. apply UCI dataset to classify the cardiovascular disease in their research [9]. The researchers from China, Yang et al. conducted a research of CVD prediction model by using random forest [10]. These different methods suit various study themes and datasets. In this essay, the author will apply Naive Bayes, Logistic Regression and Random Forest to the dataset selected and conclude that which machine learning model is most suitable and effective for this research.

This paper aims to study various potential factors that might lead to cardiovascular diseases, assess risk of cardiovascular diseases with naive bayes, logistic regression and random forest, find the relationship between variables and the diagnosis of CVD. Thus, the scientists can give valuable advises about how to prevent CVD.

In summary, after overall consideration and prediction, this paper will use the Naive Bayes, Logistic Regression and Random Forest models to study the effect of these 18 factors on CVD, i.e., whether they are factors that possibly lead to CVD.

2. Methods

2.1. Data Source

The data for this essay is collected from the Kaggle website, which was compiled by Alphiree from the online Cardiovascular Diseases Risk Prediction Dataset. This dataset cited data from 2021 Behavioral Risk Factor Surveillance System (BRFSS). This study uses and processes the 5,523 records collected as data from the BRFSS in 2021 from World Health Organization (WHO).

2.2. Variable Selection

The data for this paper count 5,523 records cited from the 2021 BRFSS totally, including those who have and do not have cardiovascular disease (Table 1). The data contains 18 factors (General Health, Checkup, Exercise, Skin Cancer, Other Cancer, Depression, Diabetes, Arthritis, Sex, Age Category, Height, Weight, BMI, Smoking History, Alcohol Consumption, Fruit Consumption, Green Vegetables Consumptions, FriedPotato Consumption).

Table 1. Logogram and numbers of the 18 factors

Elements	Logogram	Number1	CVD1
General Health	x_1	5,523	563
Checkup	x_2	5,523	563
Exercise	x_3	3,818	310
Skin Cancer	x_4	584	95
Other Cancer	x_5	589	94
Depression	x_6	1,037	133
Diabetes	x_7	968	210
Arthritis	x_8	2,127	334
Sex	x_9	5,523	563
Age Category	x_{10}	5,523	563
Height	x_{11}	5,523	563
Weight	x_{12}	5,523	563
BMI	x_{13}	5,523	563
Smoking History	x_{14}	2,255	315
Alcohol Consumption	x_{15}	5,523	563
Fruit Consumption	x_{16}	5,523	563
Green Vegetables Consumptions	x_{17}	5,523	563
Fried Potato Consumption	x_{18}	5,523	563

Number 1: The number of people suffering from the disease. CVD1: The number of cardiovascular disease patients suffering from the disease. Table 1 provides a detailed overview of the population size and the number of individuals diagnosed with CVD. As displayed in Table 1, for the sake of writing, the logogram of the variables is as indicated above. The sample of data is 5,523 individuals, of which 563 have CVD.

Table 2. Gender distribution by age group

Age	[18-24]	[25-29]	[30-34]	[35-39]	[40-44]	[45-49]	[50-54]	[55-59]	[60-64]	[65-69]	[70-74]	[75-79]	80+
Number 2	289	241	312	323	355	339	415	494	639	650	665	395	406
Female	135	139	183	186	203	205	231	269	389	388	373	236	251
Male	154	102	129	137	152	134	184	225	250	262	292	159	155
Cancer2	3	3	6	3	6	17	35	52	73	93	113	72	87

*Number2: The number of people in each age group.
**CVD2: The number of people in each age group who have CVD.

Table 2 provides a detailed breakdown of females and males, and CVD patients across various age groups. The predominant age categories of people in the data were aged 60-64 years old, 65-69 years old and 70-74 years old.

2.3. Method Introduction

In this study, three machine learning algorithms namely Naive Bayes, Logistic Regression and Random Forest were applied one after another to analyze the relationship between Heart Disease and other 18 variables. These methods were implemented using SPSSAU, and the performance of the model was assessed based on key evaluation criteria such as accuracy, precision, recall and the F1 score.

Naive Bayes is a probabilistic machine learning model that is firmly established on Bayes' Theorem. It operates under the assumption that, for a given class label, the features are conditionally independent of each other.

It works by calculating the posterior probability of each class and predicting the class with the highest probability. However, the assumption of feature independence can sometimes limit its accuracy in more complex datasets where correlations exist among features.

Logistic Regression is a statistical model employed for binary classification tasks, which estimates the probability of a binary outcome based on one or more predictor variables. It applies a logistic function to the linear combination of input features, thereby transforming the resultant value into a probability within the range of 0 and 1. It is a model popular in medial diagnosis and so many fields. Comparing to Naive Bayes, Logistic Regression model accounts for feature interactions and correlations, making it possible for datasets where features may not be independent. But Logistic Regression model may struggle with complex decision boundaries, making it less effective than more complex models in highly nonlinear datasets.

The Random Forest algorithm embodies an ensemble learning methodology, wherein during the training phase, it constructs numerous decision trees and thereafter outputs the statistical mode of the classes pertaining to classification objectives. The Random Forest model is renowned for its capacity to manage high-dimensional tasks and its resilience against overfitting, particularly in scenarios where intricate interactive relationships exist among features.

3. Results and Discussion

3.1. Model Testing

It is showed in Table 3, Table 4, Table 5 and Table 6 that as model fitting quality is judged by precision, recall, f1-score and accuracy. The precision, recall, and f1-score of the Naive Bayes model and Logistic Regression are inadequate in both training set and testing set in the situation of "Yes", which means the existence of CVD. Hence, the Naive Bayes model and Logistic Regression model are not suitable for the selected dataset and research of CVD.

Table 3. Training set model evaluation results of Naive Bayes model

Items	Precision	Recall	F1-score	Sample Size
No	92%	86%	89%	3,967
Yes	23%	37%	28%	451
Accuracy			81%	4,418

Table 4. Testing set model evaluation results of Naive Bayes model

Items	Precision	Recall	F1-score	Sample Size
No	93%	85%	89%	993
Yes	27%	46%	34%	112
Accuracy			82%	1,105

Table 5. Training set model evaluation results of Logistic Regression model

Items	Precision	Recall	F1-score	Sample Size
No	90%	100%	95%	3,967
Yes	25%	1%	1%	451
Accuracy			90%	4,418

Table 6. Testing set model evaluation results of Logistic Regression model

Items	Precision	Recall	F1-score	Sample Size
No	90%	100%	95%	993
Yes	33%	1%	2%	112
Accuracy			90%	1,105

It can be seen from Table 7 and Table 8 that as model fitting quality is judged by precision, recall, f1-score and accuracy. The Random Forest algorithm is best suitable for the selected dataset and the research of CVD among the three machine learning models and the model fitting is acceptable.

Table 7. Training set model evaluation results of Random Forest model

Items	Precision	Recall	F1-score	Sample Size
No	100%	100%	100%	3,967
Yes	100%	100%	100%	451
Accuracy			100%	4,418

Table 8. Testing set model evaluation results of Random Forest model

Items	Precision	Recall	F1-score	Sample Size
No	90%	100%	95%	993
Yes	100%	1%	2%	112
Accuracy			90%	1,105

3.2. Discussion

Figure 1 and Table 9 show different factors' importance of their contributions to the construction of the model. These factors include General Health, Checkup, Exercise, Skin Cancer, Other Cancer, Depression, Diabetes, Arthritis, Sex, Age Category, Height, Weight, BMI, Smoking History, Alcohol Consumption, Fruit Consumption, Green Vegetables Consumption and FriedPotato Consumption. There exists a total of 5,523 samples used in the Random Forest model construction.

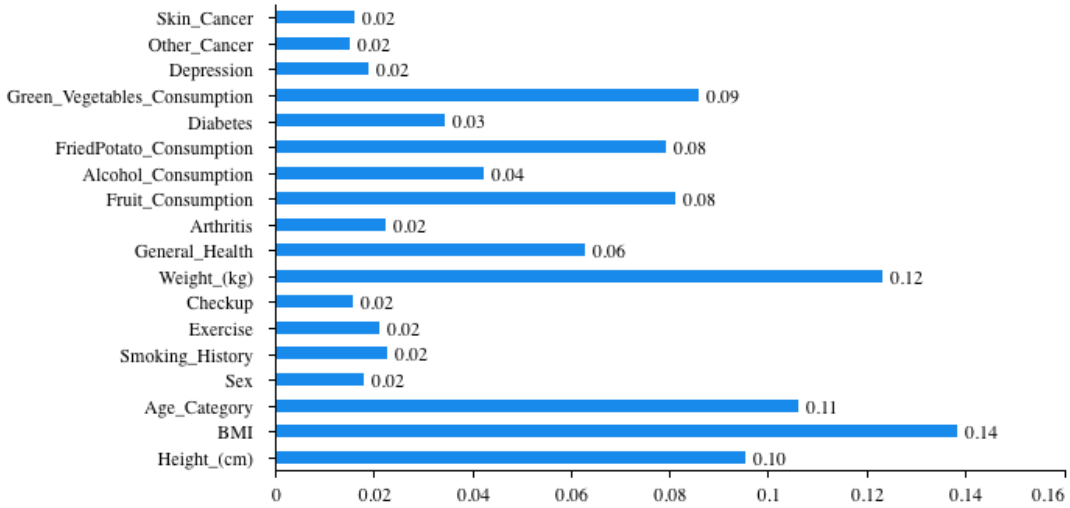

Fig. 1 Feature weight table of Random Forest model

Table 9. Feature weight table of Random Forest model

Variables	Weight
Skin Cancer	0.016
Other Cancer	0.015
Depression	0.019
Green Vegetables Consumption	0.086
Diabetes	0.035
FriedPotato Consumption	0.079
Alcohol Consumption	0.042
Fruit Consumption	0.081
Arthritis	0.022
General Health	0.063
Weight(kg)	0.123
Checkup	0.016
Exercise	0.021
Smoking History	0.023
Sex	0.018
Age Category	0.106
BMI	0.138
Height(cm)	0.096

Upon data presented in Figure 1 and Table 9, it is apparent that BMI accounts for the highest feature weight, up to 13.8%. Other Cancer has the lowest proportion at 1.5 percent. Thus, as indicated by the chart data, BMI emerges as the dominant factor, suggesting the highest hidden risk for CVD.

Based on the previous analysis, the BMI accounts for 13.84%, the Weight has a share of 12.32%, the Age Category weighs about10.62%, the Height takes up a proportion of 9.55%, the Green Vegetables Consumption, Fruit Consumption and FriedPotato Consumption account for 8.61%,

8.12%, and 7.93% respectively. The total proportion of the above seven variables is 71.01%, which holds an important position in the influencing factors of CVD.

4. Conclusion

The study selects abundant data and focuses on influencing factors that may relate to the development of cardiovascular disease. It was summarized that the having CVD may be influenced mainly by BMI, Weight, Age Category, Height, Green Vegetables Consumption, Fruit Consumption and FriedPotato Consumption, most of which have not been paid enough concentration in the past.

It is undeniable that, due to the constraint of a limited dataset, this model may encounter some errors beyond the considered factors, and the sample did not encompass all ages, ethnicities and other potentially correlated diseases, leading to possible differences, which may further impact the accuracy of the final results. However, the study retains significant value and merits. Firstly, the methods this paper using are logical and appropriate. The paper selected three classical machine learning models or algorithms namely Naive Bayes, Logistic Regression and Random Forest and conducted model tests to them respectively. The testing results shows that the Random Forest model is the most suitable one. On the one hand, when applying it in the research, a graphical approach was used to visualize the proportion of the factors' weight in the model. On the other hand, Random forest is an integrated algorithm, which specifically is a classifier comprising multiple decision trees. When compared with a single decision tree, the random forest algorithm demonstrates superior performance and can effectively mitigate overfitting. Secondly, it has some positive implications on CVD treatment.

In addition to the factors known to be associated with CVD such as Hypertension, High Cholesterol, Smoking, Diabetes, Obesity, Lifestyle and Pregnancy, more factors may be related to CVD that worth more attention and consideration, such as BMI, Age, Diet Structure, Depression, Cancer and so on. Whether these factors really associate with CVD requires future research. If more factors can be discovered, even only one more factor, it will help improving the overall well-being of patient and facilitating the progress in medicine.

References

[1] Hu Liangyuan, Bian Liu, Yan Li. Ranking sociodemographic, health behavior, prevention, and environmental factors in predicting neighborhood cardiovascular health: a Bayesian machine learning approach. Preventive medicine, 2020, 141: 106240.

[2] Fuster-Parra Pilar, et al. Bayesian network modeling: A case study of an epidemiologic system analysis of cardiovascular risk. Computer methods and programs in biomedicine, 2016, 126: 128-142.

[3] Dunn, Andrea L, et al. Reduction in cardiovascular disease risk factors: 6-month results from ProjectActive. Preventive medicine, 1997, 26: 883-892.

[4] Benschop Laura, Johannes J. Duvekot, et al. Future risk of cardiovascular disease risk factors and events in women after a hypertensive disorder of pregnancy. Heart, 2019, 105: 1273-1278.

[5] Peltola Tomi, et al. Hierarchical Bayesian Survival Analysis and Projective Covariate Selection in Cardiovascular Event Risk Prediction. BMA, 2014, 79-88.

[6] Chekouo Thierry, Sandra E. Safo. Bayesian integrative analysis and prediction with application to atherosclerosis cardiovascular disease. Biostatistics, 2023, 24: 124-139.

[7] Elsayad Alaa M, Mahmoud Fakhr. Diagnosis of cardiovascular diseases with bayesian classifiers. J. Comput. Sci., 2015, 11: 274-282.

[8] Miranda Eka, et al. Detection of cardiovascular disease risk's level for adults using naive Bayes classifier. Healthcare informatics research, 2016, 22: 196-205.

[9] Ambrish G, et al. Logistic regression technique for prediction of cardiovascular disease. Global Transitions Proceedings, 2022, 3: 127-130.

[10] Yang Li, et al. Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific reports, 2020, 10: 5245.