Predicting Cardiovascular Disease Using Simple Machine Learning Techniques

Rongrong Zhao
Living Word Shanghai, Shanghai, China
dorothyzhaorr@outlook.com

Abstract. Cardiovascular diseases (CVD) become a major health concern, which needs improved prediction models for intervention. In this study, this research explores the possibilities of using simple machine learning methods, including Logistic Regression, Decision Tree, and Random Forest, for predicting cardiovascular diseases effectively. Using a large dataset (308854 instances) containing demographic, binary, and numerical information, the research applied the above machine learning algorithms to develop predictive models. This study involved data preprocessing, including feature selection and handling duplicate values. This study then divides the dataset into two subsets: training set and testing set by 7:3 and 8:2. The Logistic Regression algorithm demonstrated good predictive performance, with an accuracy rate of 92%. Decision Tree exhibited a similar accuracy of 92%, while Random Forest underperformed both slightly with an accuracy of 91%.

Keywords: heart disease, cardiovascular disease, machine learning.

1. Introduction

Cardiovascular diseases (CVD) are disorders that affect the heart and blood vessels. These diseases include situations such as heart attacks, stroke, and heart failure. These are a popular cause of mortality worldwide and their prevention and early detection have high importance to patients.

One of the major challenges in predicting cardiovascular diseases is their complex multifactorial causes. Genetic predispositions, family history, age, gender, diabetes, obesity, smoking, and lifestyle are just some of the known factors that are easy to find. Additionally, CVD can develop silently over a long period without obvious symptoms, making diagnosis difficult.

By using the Cleveland datasets from the University of California, Irvine (UCI) machine learning repository consisting of 303 instances and 76 features, the dataset was preprocessed due to missing data, and the resulting dataset has 302 instances and 14 features.

Multiple research projects make use of the Cleveland datasets for their studies. With reference to Anitha's research (Anitha, S., 2019) presents a heart disease prediction framework in R language on how to use supervised machine learning algorithms. The study used three algorithms including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Naïve Bayes (NB). The dataset is split into training and testing sets with 70% and 30% split respectively. The result on accuracy is KNN 76.7%, SVM 77.7%, and NB 86.6%, NB shows the best result over SVM and KNN.

In a study published by Asif (Asif, A., 2021), he uses the same Cleveland dataset from UCI, which has 303 instances and 76 features, the dataset is split into training and testing sets with 80% and 20% split respectively. The study uses 11 machine learning algorithms including Logistic Regression (LR), Decision Tree (DT), SVM, Random Forest (RF), NB, KNN, AdaBoost (AdB), XG Boost (XGB), Stochastic Gradient Descent (SGD), Quadratic Discriminant Analysis (QDA) and Ensemble Voting Classifiers (EVC). The result on accuracy shows that EVC, 87%, has the best result over the other algorithms.

In Pal's study (Pal, Madhumita, 2022), the Cleveland dataset is also used, the deployed dataset has 303 instances with 13 features. However, the ratio of training and testing data for this study is not specified. Models involved are KNN and MLP with unweighted accuracy of 74% and 82% respectively.

Also, in Dalal's study (Dalal, Surjeet 2023), Jindall's study (Jindall, Harshit, 2020), and Vardhan's study (Vardhan, Valle, 2023), the Cleveland dataset is used (303 instances and 14 features). Dalal's

study deploys more advanced models while Jindall and Vardhan's studies use simpler models. Jindall's study uses K nearest neighbors (KNN), Logistic Regression and Random Forest. Vardhan's study uses DT, NB, LR, SVM and RF with a 70%/30% split on training and testing data. However, accuracy results across different models are not mentioned in the paper.

In Bhatt's study (Bhatt, Chintan, 2023), a Kaggle dataset with 70,000 instances is used. The dataset has 12 independent features, and the dataset is divided into 80% training and 20% testing data. The study involved DT, RF and Multilayer Perceptron (MLP) and XGB. The results on unweighted accuracy are 86.53 (DT), 86.92 (RF), 86.94 (MLP) and 87.02 (XGB).

These studies use datasets with small or medium number of instances, and the accuracy results range from 76%~87% where there is still room for improvement.

2. Method

This study aims to show that a dataset with a large number of instances will lower the impact of different algorithms on the accuracy of the probability of heart disease, and this is highly needed by medical professionals and patients.

The steps in this research study method are data source search that involves finding the right dataset with a large number of instances to prove the objective of this research; next in exploratory data analysis each feature is analyzed and a feature correlation heatmap is generated. To enhance the feature correlation, data engineering is deployed; then the models are developed, and run results are compared to see whether the objective is reached.

2.1. Data Source Search

This is a WHO survey named Behavioral Risk Factor Surveillance System (BRFSS) and data output for this survey in 2021 is the original sample dataset. It has the annually collected data from the which has 438,693 instances with 304 attributes. Since 1984, the BRFSS has been a national telephone survey series in the United States that gathers information from several states concerning citizens' risky behaviors, chronic illnesses, and usage of preventative services. The BRFSS dataset is then preprocessed with attributes manually selected, the refined dataset has 308,854 instances with 19 attributes and with no missing values for all 19 attributes, as mentioned in Yahaya's research (Yahaya, Lamido, 2020). This refined dataset is used for this research paper.

2.2. Exploratory Data Analysis (EDA)

In the 19 features in the refined BRFSS dataset, there are eight binary features, seven continuous numerical value features, and four categorical features. There is no missing value in the dataset. To analyze value distribution for all features in the dataset, histogram graphs are plotted for each feature. The features-specific analysis is:

- 67% of instances have good or very good health.
- 77% of instances have checkups within the past year.
- 77% of instances have claimed they have done exercise regularly.
- 8% of instances have heart disease.
- 10% of instances have skin cancer, 10% of instances have other types of cancers, 14% of instances have diabetes and 33% of instances have arthritis.
- 20% of instances have depression.
- 41% of instances have a smoking history.

The initial un-engineered features correlation heatmap is as in Fig.1. The original features from the refined dataset are lowly correlated to heart disease.

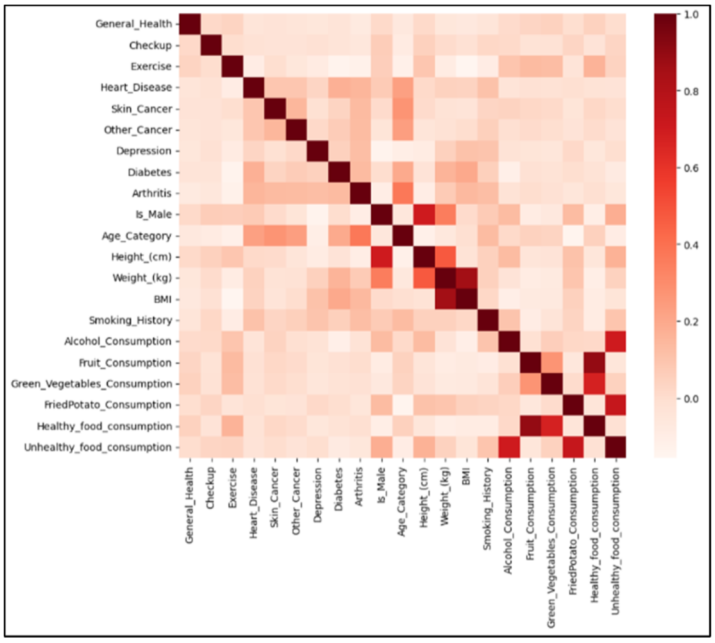

Figure 1: Correlation Heatmap Before Any Feature Engineering

2.3. Feature Engineering

To improve the model ability of the dataset, the seven numerical features are converted into categorical features with similar instances count in each category segment. The seven numerical features that are turned into categorical features are height in cm, weight in kg, BMI, consumption of alcohol (unit unknown), consumption of fruit (unit unknown), consumption of green vegetables (unit unknown) and consumption of fried potato (unit unknow).

Then, there is a second round of feature engineering to discover the impact of combining a healthy lifestyle as a group feature and an unhealthy lifestyle as another group feature:

- Healthy_Food_Consumption: adding the values of Fruit_Consumption and Green_Vegetables_Consumption together.
- Unhealthy_Food_Consumption: adding the values of Alcohol_Consumption and FriedPotato_Consumption together.

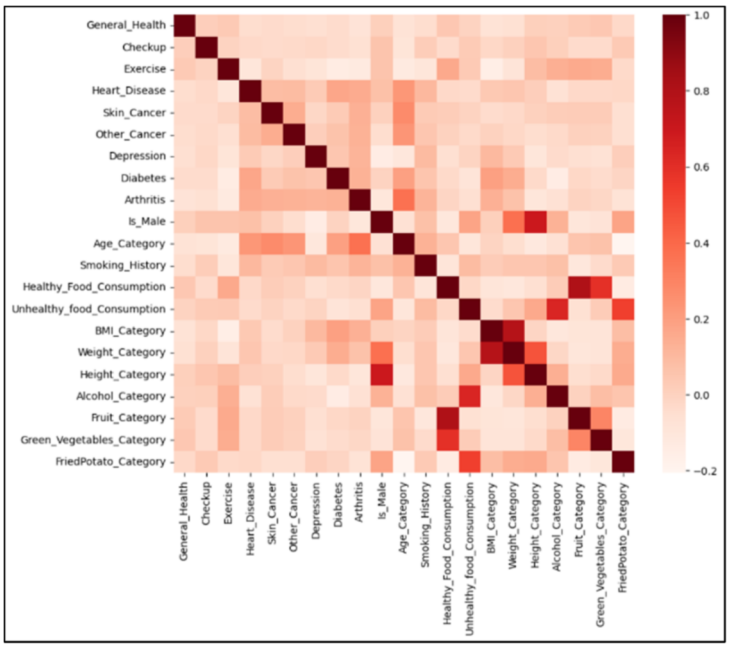

Figure 2 Correlation Heatmap After Feature Engineering

Fig.2 shows some degree of higher correlation between Heart_Disease and other features when compared with Fig.1.

3. Machine Learning Algorithms for Classification

With reference to Lupague's research (Lupague, R. M. J., 2023), three different kinds of machine learning algorithms are used in this research, namely Decision Tree, logistics regression, and Random Forest. The objectives are to test whether a bigger dataset like the BRFSS dataset (308,854 instances) versus the Cleveland dataset (304 instances) can have an impact on models built out of the same algorithm and to identify a better type of algorithm for further development.

3.1. Logistic Regression

Logistic Regression (LR) is a common statistical algorithm to handle binary classification in supervised machine learning. It predicts the probability of binary outcome based on participating features. It is selected as this algorithm gives binary outcomes even if the features are continuous or categorical independent variables. There are some assumptions when LR is being used, the relationship between independent variables is linear, independent variables are independent of each other and the outcome variable is binary. In the study, there are 20 independent features participating and the binary outcome variable is "Heart_Disease" which has values of 0 or 1.

3.2. Decision Tree

According to Scikit-Learn, Decision Tree is a supervised machine learning method without much use of parameters. This method is common for classifying and regression problems. The aim of this method is to provide simple decision rules engineering from dataset features during EDA, and those engineered feature should be corelated to the outcome feature, "Heart_Disease" in this study.

Some advantages of Decision Trees are:

- Simplicity of use and little data preparation is required.
- DT can process both numerical and categorical feature values.
- Logics are transparent and explainable by Boolean logic.
- It is possible to calculate the reliability of the DT model using statistical methods.
- Good performance can be obtained even if its assumptions are not in line with the real-life model.

However, the Scikit-learn package for model development in this research does not support categorical features, so this research only correlated to numerical features.

3.3. Random Forest

Random Forest is a widely used machine learning method. It joins the output of several Decision Trees to reach a single result. This algorithm is comprehensible and flexible. It processes both classification and regression modeling.

Some advantages of Random Forest are:

- Reduced risk of overfitting as there is a specified number of levels of depth.
- Provides flexibility: Feature-bagging makes the algorithm a practical tool for approximating missing values while maintaining its accuracy when certain amount of the dataset is not available.
- Simple to assess feature importance: simple to determine variable importance or contribution to the model.

However, there exist some challenges for the Random Forest method. The most obvious one should be slow when compared with other algorithms in processing data as it computes data for each Decision Tree. Since it takes in a large dataset, it requires more resources to store the dataset too.

4. Machine Learning Algorithms for Classification

With reference to Aljanabi's study (Aljanabi, Maryam, 2018), Accuracy, recall, and f1-score are the metrics employed in this investigation. To help understand these metrics without further research on their detailed definition, a short explanation of these metrics is:

- Accuracy: it represents the percentage of accurate results over the total number of instances in the run
- Precision: it displays the proportion of genuine positives over the total of both true and false positives.
- Recall: it is the proportion of genuine positives to the total of both true positives and false negatives.
- F1-Score: it is calculated using formula (1) to represent a balanced mean value of precision and recall.

$$\text{F1-Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision+ Recall}} \quad (1)$$

Different ratios of testing data are used in this study; it is shown that the number of instances is large enough to be independent to affect the performances of algorithms. The results shown below for individual algorithms are for 20% testing data and 30% testing data respectively.

4.1. Individual Algorithm Results

The result for the *Logistic Regression* algorithm is as below in Table 1:

Table 1. Run Results of LR, DT and RF with 20% and 30% Testing Data

	20% Testing Data				30% Testing Data			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-score	Support
Logistic Regression								
"0"	0.92	1.00	0.96	56677	0.92	1.00	0.96	85100
"1"	0.54	0.03	0.06	5078	0.53	0.03	0.06	7533
accuracy			0.92	61755			0.92	92633
macroavg	0.73	0.52	0.51	61755	0.72	0.51	0.51	92633
weightedavg	0.89	0.92	0.88	61755	0.89	0.92	0.88	92633
Decision Tree (Five Levels)								
0	0.92	1.00	0.96	56677	0.92	1.00	0.96	85100
1	0.60	0.02	0.03	5078	0.59	0.02	0.04	7533
accuracy			0.92	61755			0.92	92633
macroavg	0.76	0.51	0.49	61755	0.75	0.51	0.50	92633
weightedavg	0.89	0.92	0.88	61755	0.89	0.92	0.88	92633
Random Forest								
0	0.92	0.99	0.95	56677	0.92	0.99	0.96	85100
1	0.34	0.06	0.10	5078	0.34	0.05	0.09	7533
accuracy			0.91	61755			0.91	92633
macroavg	0.63	0.52	0.53	61755	0.72	0.52	0.52	92633
weightedavg	0.87	0.91	0.88	61755	0.89	0.91	0.89	92633

4.2. Comparison

The below table is the comparison for 20% and 30% testing data and for different algorithms.

Table 2: Comparison of Model Results Between 20% & 30% Testing Data-Part 1

	Accuracy (%)		Precision		Recall	
Algorithm	20% Test Data	30% Test Data	20% Test Data	30% Test Data	20% Test Data	30% Test Data
LR	92%	92%	0.92	0.92	1.00	1.00
DT	92%	92%	0.92	0.92	1.00	1.00
RF	91%	91%	0.92	0.92	0.99	0.99

Table 3: Comparison of Model Results Between 20% & 30% Testing Data-Part 2

	Accuracy (%)		Precision of Weighted Average		Recall of Weighted Average		F1 Score of Weighted Average	
Algorithm	20% TestData	30% TestData	20% TestData	30% TestData	20% TestData	30% TestData	20% TestData	30% TestData
LR	92%	92%	0.89	0.89	0.92	0.92	0.88	0.88
DT	92%	92%	0.89	0.89	0.92	0.92	0.88	0.88
RF	91%	91%	0.87	0.88	0.91	0.91	0.88	0.89

It shows that all three types of models have similar accuracy results, which means all of these simple classification algorithms are similar. However, it is worth noting that all three models have Precision much lower than Recall, no matter for 20% testing data run or 30% testing data run.

5. Summary

Cardiovascular disease is still high impact and high occurrence major disease worldwide. The urgency to better predict, pre-intervention, and deployment of machine learning models to front-line healthcare institutes is easy to understand. It is obvious that the refined BRFSS dataset has overcome the issue of a low number of instances that the Cleveland dataset has. Feature re-selection from the original data set, feature reengineering, and more different types of statistical-based algorithms should be tried to overcome the issue of lower precision and improve accuracy.

References

[1] Aljanabi, Maryam, Hijjawi and Qutqut. (2018) Machine Learning Classification Techniques for Heart Disease Prediction: A Review. *International Journal of Engineering and Technology*, October 2018.

[2] Anitha, S., & Sridevi, N. (2019). Heart Disease Prediction Using Data Mining Techniques. *Journal of Analysis and Computation* (2), 48-55.

[3] Asif, A., Nishat, Mirza & Faisal, Fahim (2021) Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease. *Engineering Letters* 29 May 2021 (2) pp.731-741

[4] Bhatt, Chintan M., Patel, Ghetia and Pier. (2023) Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms,* 2023, 16, 88.

[5] Dalal, Surjeet, Goel, Onyema, Alharbi, Mahmoud, Algarni and Awal. (2023) Application of Machine Learning for Cardiovascular Disease Risk Prediction. *Computational Intelligence and Neuroscience,* Volume 2023.

[6] Jindall, Harshit, Agrawal1, Khera1, Jain and Nagrath. (2020) Heart Disease Prediction Using Machine Learning Algorithms. *IOP Conf. Series: Materials Science and Engineering*, 1022 (2021) 012072.

[7] Lupague, R. M. J., Mabborang, R. C., Bansil, Alvin & 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), p44-58, 2023.

[8] Pal, Madhumita, Parija, Panda, Dhama and Mohapatra. (2022) Risk Prediction of Cardiovascular Disease Using Machine Learning Classifiers. *Open Medicine*, 2022; 17: 1100–1113.

[9] Vardhan, Valle, Kumar, Vardhini, Varalakshmi and Kumar. (2023) Heart Disease Prediction Using Machine Learning. *Journal of Engineering Sciences*, Vol 14 Issue 04,2023

[10] Yahaya, Lamido, Oye, Nathaniel David & Garba, Etemi Joshua. (2020) A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques. *American Journal of Artificial Intelligence*, Vol. 4. NO. 1 20202. pp. 20-20.