Cardiovascular Diseases Risk Prediction Based on Machine Learning Models

Steven Kalung Li [1, *], Yuli Peng [2]

[1] Wellington College, Crowthorne, United Kingdom

[2] Chongqing Foreign Language School, Chongqing, China

* Corresponding Author Email: Lis@wellingtoncollege.org.uk

Abstract. Cardiovascular Diseases (CVD) are the leading cause of death worldwide, CVD has a high death rate of 95% as it causes a blockage in blood vessels, which prevents blood from flowing to the brain or heart. Necessitating accurate risk prediction for effective prevention and management becomes a necessity of lowering death caused by cardiovascular disease. This paper presents a comprehensive study utilizing Logistic Regression and Decision Tree models implemented via Python to predict CVD risk. The models were trained on a dataset comprising numerous variables such as age, checkup, sex, and BMI which are essential factors and indicators that are related to health and wellbeing. A detailed evaluation would be produced, including accuracy, precision, and confusion matrix, demonstrates the efficiency of the proposed approach, with the Logistic Regression model achieving an accuracy of 91.94% and a precision of 71.94%. followed by the decision tree model achieving an accuracy of 86.14% and a precision of 19.7%.

Keywords: Machine learning, CVD, logistic regression, decision tree.

1. Introduction

Cardiovascular Diseases (CVD) are perceived as a great threat to global health, contributing to 17.9 million deaths annually [1]. Early prediction and diagnosis of CVD are critical in initiating timely preventive measures and treatment. Traditional risk prediction methods often fall short in their predictive power, leading to the need for more precise and reliable techniques. In the era of big data and machine learning, the application of computational models developed with Python leads to a promising direction of research. The project focuses on the application of two models, Logistic regression and decision tree, to predict CDV risk. The aim of the project includes.

- Exploring the relationship of risk factors to CDV
- Use and compare Logistic regression and Decision tree models.
- Evaluate the performance of the models based on results such as accuracy and precision.

The remainder of the paper is structured as follows: Section 2 focuses on the review of previous research done. Section 3 describes the data and methodology; Section 4 presents the results Section 5 concludes the paper with implications and future directions.

2. Literature review

2.1. Traditional risk prediction models

Initially, CVD risk assessment was performed using statistical methods and scoring systems like the Framingham Risk Score. These models, while pioneering, have shown limitations in their predictive accuracy, particularly in diverse populations [2]

2.2. Logistic regression in CDV prediction

Logistic Regression has been a popular choice in medical research for binary classification problems such as CVD risk prediction. Recent study conducted by Lakshmanarao, A., Y. Swathi, and P. Sri Sai Sundareswar. [3] attempted to use the logistic regression model to try to achieve the goal

of predicting CDV. The results show a precision of 68% and 67.5% in accuracy. The low performance could be caused by factors like small data sets or inefficient ways of splitting the data set.

2.3. Decision tree in CVD prediction

Decision Tree models, such as the one used by Pal, Madhumita, and Smita Parija [4], have also gained attention for their interpretability and ability to handle non-linear relationships. They have been applied to various healthcare domains, including CVD prediction, offering an alternative to traditional statistical methods. Their results of an accuracy of 86.9% show that they have achieved higher accuracy than the previous paper. But this has also increased as there were insufficient testing taking place as there was only 61 data used for testing. This is shown in their confusion matrix.

2.4. Comparison between studies

Several studies have conducted comparisons between different machine learning models, including Logistic Regression and Decision Trees, for predicting CVD risk. The findings have varied, with studies of Ambrish, G favoring Logistic Regression for its simplicity and achieving an accuracy of 82.56% [5] and others like Pasha, Syed Nawaz Pasha, Syed Nawaz highlighting the flexibility of Decision Trees which achieved an accuracy of 81.97[6].

3. Methodology

The main steps within this model training are 1) Data cleaning 2) EDA 3) Data scaling 4) Data splitting 5) Model training 6) Model testing 7) Results evaluation.

3.1. Data cleaning

Firstly, we used the isnull function to achieve the checking of null values which could cause errors in later steps such as training data and splitting data as shown in Figure 1.

The range above checking include the following factors: General Health checkup, Exercise, Heart Disease, Skin Cancer, Other Cancer, Depression, Depression, Diabetes, Arthritis, Sex, Age_Category, Height_(cm), Weight_(kg), BMI, Smoking History, Alcohol Consumption, Fruit Consumption, and Green_Vegetables_Consumption, FriedPotato_Consumption type.

```
General_Health                0
Checkup                       0
Exercise                      0
Heart_Disease                 0
Skin_Cancer                   0
Other_Cancer                  0
Depression                    0
Diabetes                      0
Arthritis                     0
Sex                           0
Age_Category                  0
Height_(cm)                   0
Weight_(kg)                   0
BMI                           0
Smoking_History               0
Alcohol_Consumption           0
Fruit_Consumption             0
Green_Vegetables_Consumption  0
FriedPotato_Consumption       0
dtype: int64
```

Figure 1. Null variable output (Photo/Picture credit: Original)

After checking data, data mapping from strings to integers was used as strings could not be used as a variable when training data.

3.2. EDA

Where we used a few graphs to show the relationship between a few important variables compared to the rate or percentage of having CVD. This process involves using functions such as Seaborn library and matplotlib library.

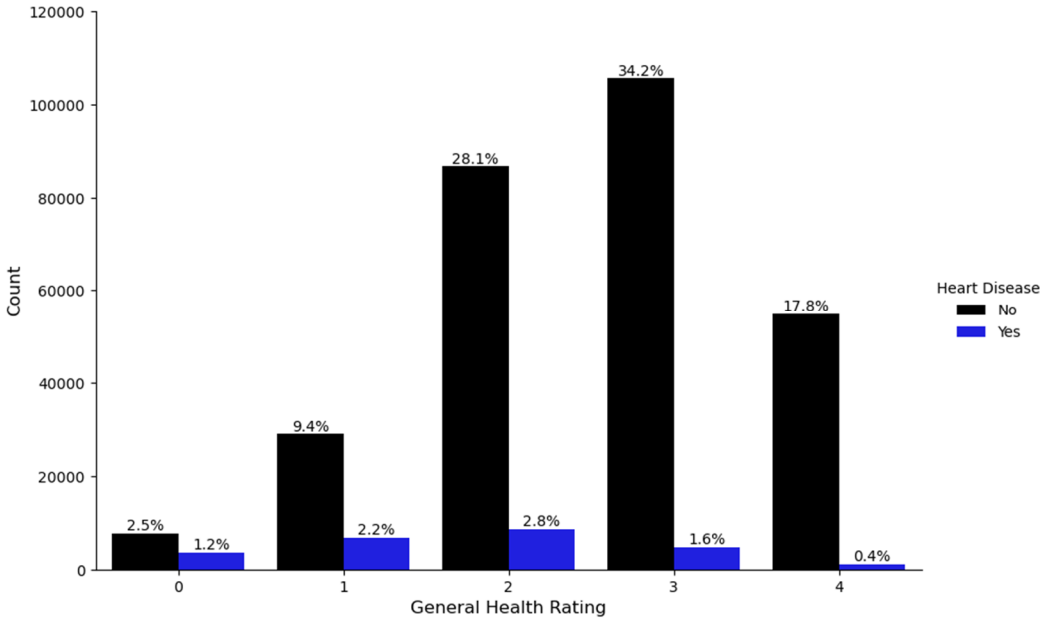


Figure 2. General Health Rating (Photo/Picture credit: Original)

This Figure 2 uses the health rating that has been mapped to poor -1, fair -1, good -2, very good -3, and excellent -4. Where this graph shows the lower the general health rating is the chance of having CVD increases.

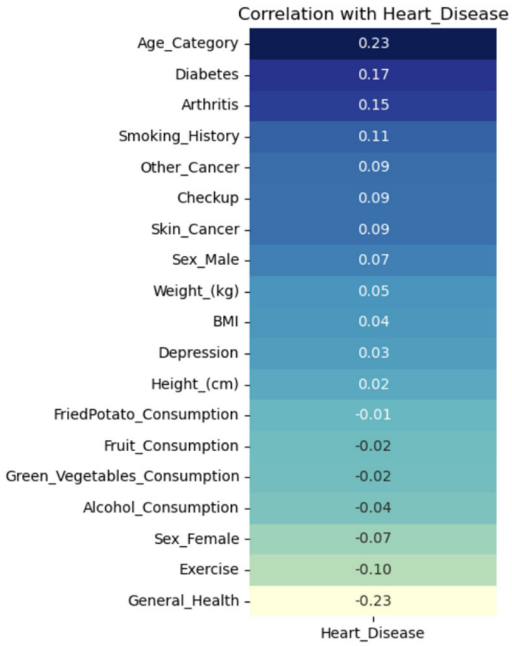


Figure 3. The correlation between factors and Heart_Disease (Photo/Picture credit: Original)

This heat map Figure 3 shows the correlations of different variables compared with heart disease to show which variables contribute and are most important to predicting heart disease.

3.3. Data scaling

Standard scaling is a common preprocessing technique used in statistical and machine learning analyses. It's used to standardize the features of our dataset, i.e., to rescale the variables so that they have a mean of 0 and a standard deviation of 1. This can be particularly important when your features have different units or different scales, as many perform better when the input features are on the same scale.

3.4. Splitting data

The model would need to be trained and tested, meaning the data would also need to be split into 2, in our case we allocated 80% of data for training and 20% of data for testing.

3.5. Models training/testing

3.5.1 Logistic regression [7]

The core of Logistic Regression shown in equation (1) and equation (2) is the logistic function above where the results would always be 0 or 1 allowing it suitable for representing a probability.

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

$$Y_i = \beta_0 + \beta_1 X_i \tag{2}$$

Where X1 represents the feature variables and B0, B1 represents the Coefficients that need to be learned from the data.

1) We set “model” as our variable for the Logistic regression function and made a parameter of 200 iterations max. as this would be the most optimized parameter for this dataset.
2) Then we trained the model using the scaled training sets created in splitting data.
3) Finally, use the function “predict” to predict and get a final outcome of the performance of the model.

3.5.2 Decision tree [8]

1) At every node in the tree, the algorithm selects the best feature to split the data.
2) Once the best feature is chosen, the data is split into subsets based on that feature.
3) After splitting the data, the algorithm creates child nodes for each subset.
4) The algorithm continues to split the data, create child nodes, and repeat the process recursively until one of the stopping conditions is met.
5) Once the tree is built, the leaves represent the final predictions.
6) To make a prediction for a new data point, the decision tree "routes" the data point down the tree, following the branches that correspond to the data point's features. Once it reaches a leaf, the prediction associated with that leaf is returned.

4. Results and discussion

4.1. Logistic regression model

The model has shown a high accuracy of 91.94% with a lower precision score of 71.94%. where this a caused by a high TN of 56470, where the model correctly predicted patients without heart disease. Followed by a low FP of 304, which is when the model incorrectly predicts patients with heart disease.

4.2. Decision tree model

The results show a reasonable degree of accuracy of 86.14 but a low precision of only 19.7. This is caused by the high false positives of 4724 and false negatives of 3838.

4.3. Comparison between the models

The model with the better performance is the Logistic regression model, as its accuracy is 5.8% and its precision is 52.24% higher. The large increase in precision is caused by its lower false positive. A false positive would cause a high consequence of unneeded stress and a high cost of unneeded health checks.

4.4. Comparison with other papers

In our project, the Logistic regression model achieved an accuracy of 91.94 when compared with Ambrish, G's result [9] of 84.85% has shown a large increase. This is caused by not overfitting the data set and using a sensible number of variables.

The second paper from Pal, Madhumita, and Smita Parija [10] has a lower accuracy of 88.4% than our model because our model has better scaling. However, they have higher precision as they have divided their data better.

5. Conclusion

In our exploration of predicting CVD, we utilized machine learning techniques like Logistic Regression and Decision Trees. With careful consideration of features such as age, blood pressure, cholesterol levels, and lifestyle factors, we were able to create models to gauge the risk of CVD.

Our Logistic Regression model achieved an impressive 91.94% accuracy. This success underscores the model's capability to generalize well to unseen data, providing a valuable tool for understanding and predicting CVD risks. The use of standard scaling helped to maintain consistency in our data, allowing for better convergence during the training process.

Future steps could include: 1) Exploring Other Algorithms: We can investigate additional machine learning techniques to see if they offer improvements or insights. 2) Ethical Considerations: As we work with medical data, understanding and adhering to privacy concerns is essential. 3) Community Collaboration: Collaborating with local healthcare providers might allow us to test our model's real-world applicability.

Authors Contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] World Health Organization Cardiovascular diseases, 2023, https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] Lloyd-Jones, D M et al. Framingham risk score and prediction of lifetime risk for coronary heart disease. The American Journal of Cardiology 94.1 2004, 20-24

[3] Lakshmanarao, A Y et al. Machine learning techniques for heart disease prediction. Forest 95.99, 2019, 97.

[4] Pal, M, and Smita P. Prediction of heart diseases using random forest. Journal of Physics: Conference Series. Vol. 1817. No. 1. IOP Publishing, 2021.

[5] Ambrish, G, et al. Logistic regression technique for prediction of cardiovascular disease." Global Transitions Proceedings 3.1, 2022, 127-130.

[6] Pasha, S N, et al. Cardiovascular disease prediction using deep learning techniques. IOP conference series: materials science and engineering. Vol. 981. No. 2. IOP Publishing, 2020.

[7] LaValley M P. Logistic regression. Circulation, 2008, 117(18): 2395-2399.

[8] Myles A J, Feudale R N, Liu Y, et al. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 2004, 18(6): 275-285.

[9] Ambrish, G., et al. Logistic regression technique for prediction of cardiovascular disease. Global Transitions Proceedings 3.1, 2022, 127-130.

[10] Pal, M, and Smita P. Prediction of heart diseases using random forest. Journal of Physics: Conference Series. Vol. 1817. No. 1. IOP Publishing, 2021.