Comprehensive Analysis of Factors Influencing Heart Disease Risk

Qi Wu

Institute of Science and Technology, Beijing Normal University-Hong Kong Baptist University Joint International College, Zhuhai, China

s230026165@mail.uic.edu.cn

Abstract. Research Background and Significance: Heart disease continues to be a leading cause of mortality globally, posing significant challenges in the realms of prevention and management. The complexity of cardiovascular diseases arises from an interplay of genetic, lifestyle, and environmental factors, making their study and prediction critically important for public health. The integration of machine learning techniques into medical research has opened new avenues for understanding these diseases, significantly advancing the capabilities of predictive models. This paper leverages a comprehensive dataset from Kaggle, incorporating a diverse range of variables such as lifestyle habits and physiological markers known to influence cardiovascular health, which provides a foundation for robust analytical exploration. **Contributions of This Paper:** This study makes several significant contributions to the field of cardiovascular disease research. Firstly, it employs advanced statistical techniques such as Principal Component Analysis (PCA) and K-means clustering to effectively reduce data multicollinearity and dimensionality, which enhances the clarity and reliability of the findings. The PCA approach successfully condensed the data into principal components that explain a substantial portion of the variability, while K-means clustering categorized the data into meaningful risk profiles. Secondly, this paper demonstrates the utility of factor analysis in identifying major risk factors like smoking, age, and gender, furthering the understanding of their roles in heart disease risk. Finally, the application of various machine learning models.

Keywords: Heart disease; Comprehensive analysis; Factors influencing.

1. Introduction

The heart plays a crucial role in the detoxification process by returning metabolic waste products to the liver and kidneys for processing, thus preventing the buildup of harmful substances that could lead to severe health issues [1]. The etiology of cardiovascular disease is multifaceted, involving a combination of hereditary factors, lifestyle choices, and environmental influences. Despite the myriad risk factors contributing to cardiovascular disease, one of the most pressing challenges in contemporary medicine is the development of effective preventative strategies [2]. Recently, machine learning has proven to be an invaluable tool in medical research, providing significant insights into disease prediction. This paper utilizes a comprehensive dataset from Kaggle to model the risk factors associated with heart disease [3]. The dataset includes a wide range of attributes, such as lifestyle habits and physiological markers, all known to influence the development of cardiovascular conditions. The subsequent analysis aims to train robust models and generate predictions consistent with initial findings [4], thereby enhancing the decision-making capabilities of healthcare professionals.

In preparing the data, duplicate entries were removed, and missing values were addressed. "Blood pressure" was categorized into "diastolic" and "systolic" to structure the dataset effectively, which was then standardized to include information on 8763 patients across 27 variables related to heart health [5]. This work outlines the process of identifying and analyzing various factors that influence heart disease risk by employing advanced statistical techniques like Principal Component Analysis (PCA) and K-means clustering. These methods are used to reduce multicollinearity and the dimensionality of the data, which encompasses environmental, lifestyle, and genetic factors affecting heart disease. The objective is to understand the complex interactions of these variables and to develop predictive models that provide actionable insights for medical practitioners. The PCA

effectively condensed the dataset into 21 principal components that explain 84.9% of the variability, while K-means clustering grouped the data into three distinct risk profiles. Factor analysis was also conducted to further explore and identify key factors such as smoking, age, and gender, which play significant roles in heart disease risk. The final models, including Decision Trees, Random Forest, Logistic Regression, and K-Nearest Neighbors, were evaluated through cross-validation to prevent overfitting, with the Decision Tree model demonstrating the highest accuracy. These findings underscore the critical predictors of heart disease and offer valuable insights for targeted prevention strategies.

2. Advanced Data Reduction and Clustering Techniques

2.1. Principal Component Analysis

Principal component analysis (PCA) is a statistical technique that changes a collection of variable components that may be linearly linked into a set of variable components that are linearly uncorrelated. This transformation is accomplished via the use of orthogonal transformation [6]. A significant reduction in the impact of multicollinearity is achieved by the use of principle components, which are the new variables. The order in which the principal components are ranked will be determined by the amount of variation that can be explained by the dataset.

Through the use of principal component analysis, the elements in this data set have been broken down into 21 main components. Out of these 21 principal components, 17 of them are able to explain 84.9% of the data itself [7]. The future study will make use of these seventeen principal components, which will allow for the dimensionality reduction of variables to be achieved while preserving crucial information and successfully reducing the effects of multicollinearity on the findings. As can be seen in It is clear from the core data point set that a large number of people have traits in both of these segments, whereas those individuals who are located a significant distance from the central point are those who have qualities that are independent of one another. As shown in Fig 1.

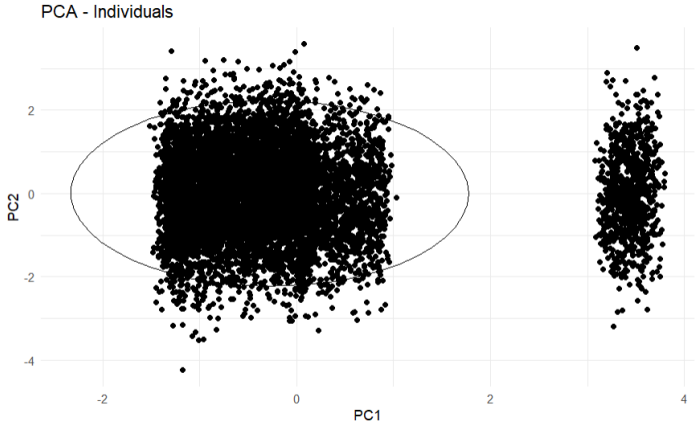

Fig. 1 PCA-Individuals Scatter Plot (Photo credit: Original).

The information shown in Fig. 2 reveals that the colour scale is a representation of the contribution level, and the progression from red to green denotes the contribution level from high to low. This can be recognised by reading the material. The elements that have the greatest influence on the first principal component are and, respectively, and have a significant influence on the second principal component, which may also be used to depict the main characteristics in the initial data in order to comprehend how they contribute to the risk of heart disease.

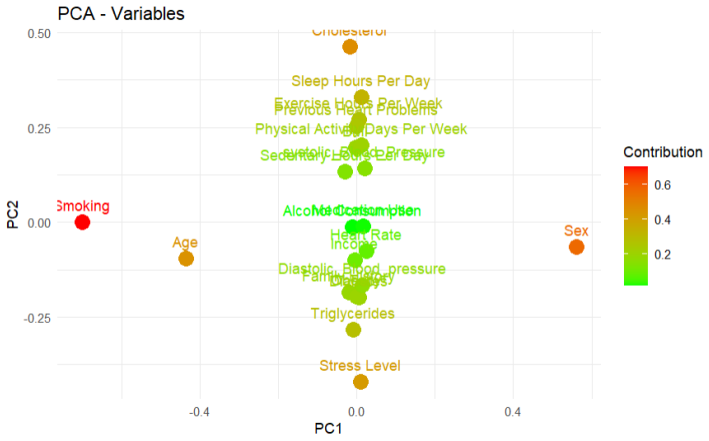

Fig. 2 PCA-Variables Contribution (Photo credit: Original).

2.2. K-Means Clustering Algorithm

The K-Means clustering method represents an accessible and effective approach to clustering analysis. This is achieved by calculating the intra-cluster distances with the lowest sum of squared values. The primary objective of this technique is to partition a set of items into a specified number of clusters, based on the characteristics exhibited by each individual object. Consequently, the elements within the cluster will exhibit a high degree of similarity, whereas those outside the cluster will display a low degree of similarity. Consequently, the resulting clusters will be among the most effective possible.

In order to determine the optimal number of divisions to employ, the elbow rule and the contour coefficient rule were used in combination with one another. In accordance with the elbow rule, the gradient of the function graph line would become gradual when the number reached three. However, the contour coefficient method indicated that the mean silhouette width would reach its maximum at the same number. Both hypotheses were predicated on the assumption that the optimal number would be three. The results presented in Fig. 3 and 4 demonstrated that the optimal number of clusters was indeed three.

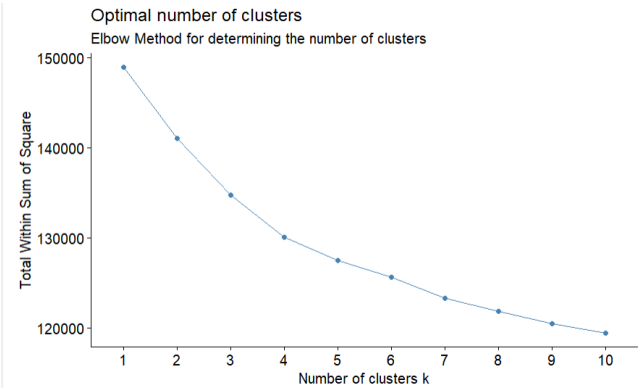

Fig. 3 Elbow Method for Optimal Clusters (Photo credit: Original).

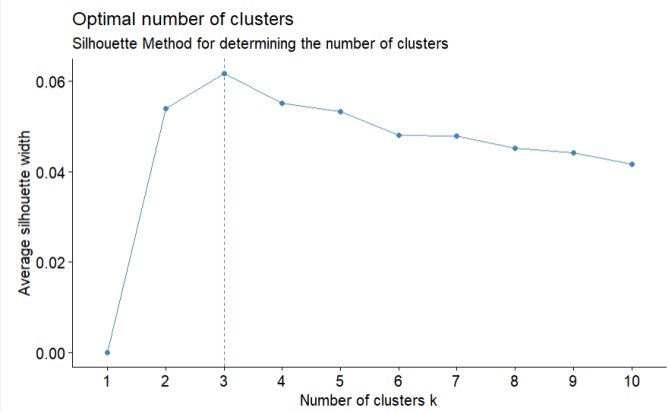

Fig. 4 Silhouette Method for Optimal Clusters (Photo credit: Original).

This is done in order to complete the clustering process. The output of the clustering process comprises three distinct risk characteristics, each represented by a separate cluster. Subsequently, the cross-validation modelling technique will ensure that the data are distributed in a fair and equal manner.

2.3. Factor Analysis

The application of factor analysis enables the examination of the intrinsic relationships between variables. This suggests that a minimal number of variables, which are not inherently related to one another, may be employed to describe the characteristics of a considerably larger number of variables. Once the factors have been extracted, it is possible to ascertain the loading of each variable on each factor. This is an alternative designation for the correlation identified between the factor and the variable.

Once the examination of the key components was complete, the data were employed as the foundation for the analysis of the factors. As illustrated in Fig. 5, the data indicate the presence of three latent components. Two factors were included in the initial latent factor. These were "smoking" and "gender," with correlation values of 0.9 and -0.6, respectively. Both of these factors exerted a considerable influence on the principal component. The principal component was significantly influenced by each of these variables to a considerable extent. Included in the second latent component, with a coefficient of 0.7, was variable "age" which was subjected to analysis.

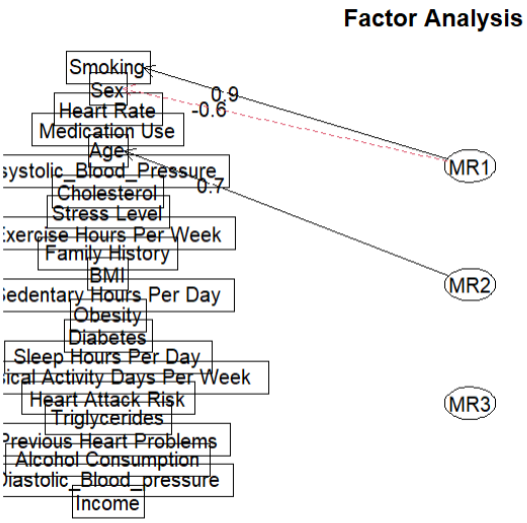

Fig. 5 Variable Importance in Factor Analysis (Photo credit: Original).

3. Establishing a Heart Disease Risk Prediction Model

3.1. Cross-Validation

In addition to this, it helps to ensure that the model generalizes correctly and avoids overfitting from occurring. Each time during the iteration process, the training set is comprised of k-1 subsets, while the remaining subset is used as the test set. k times, with each iteration utilizing a different subset as the test set, the operation is continued until it reaches its conclusion. Next, the calculation is made to get the average of the results acquired from all k iterations.

k is set to ten in this modeling, the principal components of the completed clustering are set as a new data set, twenty percent is used as the test set, and eighty percent of the training set and twenty percent of the validation set are set in the training set to ensure that the data are random and equal. Lastly, the test set is set to twenty percent. In addition, a number of performance indicators applicable to the training set are shown here, including the F1 score and the sensitivity metrics.

3.2. Random Forest Model

The Random Forest algorithm is well recognized as an outstanding method of classification for supervised learning, with strong capabilities for carrying out regression tasks in addition. The system uses many decision trees to generate predictions, with each tree giving a vote toward the final expected classification [8].

3.3. Decision Tree Classifier

With its hierarchical structure, a decision tree is a representation of the mapping connection that exists between the characteristics that are input and the targets that are produced. This provides a straightforward explanation as well as a speedy training experience.

3.4. Logistic Regression

The logistic regression model demonstrates the highest accuracy when dealing with a dichotomous categorical response variable. Within the realm of machine learning algorithms, the logistic regression model is commonly employed for classification purposes. It operates on the principle of likelihood by assigning observations to discrete classes through logistic regression analysis [9].

3.5. k-Nearest Neighbor Model

In order to create a prediction, the k-Nearest Neighbour Model is a straightforward classification and regression technique. It does this by comparing the degree of similarity between a new sample and the samples that are included in the training set. It then chooses the k samples that are closest to the new sample and calculates the mean or median of the output values of these k samples. Finally, it determines the value that is anticipated to be associated with the new sample when it is applied to the training set. Its predictive accuracy may suffer as a result of its reliance on extensive datasets and extensive trees, which may compromise the transparency of the generated outcomes [10].

4. Results and Discussion

Table 1. Shows the accuracy rates or AUC values of the four models with or without validation sets.

	Logical Regression	Random Forest	K Nearest Neighbour	Decision Tree Classifier
train_ y with validation set	AUC: 0.5019 Accuracy: 0.6391		AUC: 0.4949 Accuracy: 0.5711	AUC: 0.4890 Accuracy: 0.6202
train_ y without validation set	AUC: 0.5056 Accuracy: 0.6391	Accuracy: 0.6345	AUC: 0.4949 Accuracy: 0.5711	AUC: 0.4890 Accuracy: 0.6202
test_ y with validation set	AUC: 0.5103 Accuracy: 0.6391		AUC: 0.5888 Accuracy: 0.5888	AUC: 0.6476 Accuracy: 0.6476
test_y without validation set	AUC: 0.5103 Accuracy: 0.6391	Accuracy: 0.6345	AUC: 0.5888 Accuracy: 0.5888	AUC: 0.6476 Accuracy: 0.6476
val_ y	AUC: 0.5139		AUC: 0.6052	AUC: 0.6452

According to the facts that are shown in the table 1, it is feasible to make the observation that the decision tree model has an excellent degree of fitting capacity. This is something that can be seen. The preliminary hypothesis that was created before to the creation of the model estimated an accuracy rate of seventy percent. This was done before the framework was constructed. On the other hand, when put into reality, the ideal model only achieves an accurate rate of 64.76%, which implies that there is a chance of underfitting. For the second test, we decided to skip the verification set approach and instead divide our dataset into two parts. This choice was taken because we wanted to ensure that the results of the second test were accurate. The accuracy rate that was reached by this division was the highest that was possible to achieve to that point.

For another way of putting it, the accuracy rate is the probability that a certain model that makes use of data is able to accurately predict whether or not there is a risk of getting heart disease. Nevertheless, in spite of the fact that we have implemented this method, we have not yet been able to achieve the level of success that we had anticipated in terms of the accuracy rates.

When dealing with decision tree models, it is essential to bear in mind that adjusting parameters such as maximum depth may have an impact on the accuracy of predictions. This is something that should be kept in mind consistently. It is likely that a lower depth may result in an oversimplification of the complex relationships that are present within the data; yet, it will help to prevent overfitting from occurring. More depths, on the other hand, may lead to models that are too intricate and prone to overfitting issues rather than being straightforward. Therefore, in order to get the highest possible levels of accuracy during the course of the experiment, modifications were made to the parameters in order to achieve the greatest possible depth value rises. The adjustments did not result in significant improvements to the general accuracy, despite the fact that this was the case.

Additionally, the F1 scores that were produced from the training sets via the use of the K-fold cross-validation techniques indicate a balanced performance that is below 80%, which is considered to be relatively unsatisfactory. Despite the fact that the sensitivity is still rather high, namely more than 90 percent, this demonstrates that our approach is able to correctly detect true positive events. This is a more favorable interpretation of the scenario.

5. Conclusion

During the course of the research, it was discovered that the chance of having a heart attack is significantly influenced by a number of variables, including age, gender, smoking habits, stress levels, and cholesterol levels, among others. Some of these factors are listed below. When applied to a larger population, predictive models have the ability to provide services that are personalized to a certain

degree, including risk assessment and health management. This is because it is considered to be more helpful. With the addition of real patient data or the use of more sophisticated models, the effectiveness of predictive models may now be enhanced, which will contribute to the decrease of rates of cardiovascular disease and mortality while simultaneously boosting the delivery of healthcare that is delivered.

References

[1] Mishra A., Wang Y., Zhou Q., et al. Retracted: Implementation of a Heart Disease Risk Prediction Model Using Machine Learning. Computational and Mathematical Methods in Medicine, 2023.

[2] Turabieh H. A Hybrid ANN-GWO Algorithm for Prediction of Heart Disease. American Journal of Operations Research, 2016, 6: 136-146.

[3] Wang R., Zhu J., Wang S., Wang T., Huang J., Zhu X. Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking. In: International Journal of Multimedia Information Retrieval, 2024, 13(4): 39.

[4] Asgarov E. A Comprehensive Analysis of Machine Learning Techniques for Heart Disease Prediction. Open Access Library Journal, 2024, 11: 1-17.

[5] Fatima M., Pasha M. Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications, 2017, 9: 1-16.

[6] Zhu, X., Guo, C., Feng, H., Huang, Y., Feng, Y., Wang, X., Wang, R.: 'A Review of Key Technologies for Emotion Analysis Using Multimodal Information', Cognitive Computation, 2024, 1, (1), pp. 1-27.

[7] Yusoff M.I.M. Machine Learning: An Overview. Open Journal of Modelling and Simulation, 2024, 12: 89-99.

[8] Gabriel J. A Machine Learning-Based Web Application for Heart Disease Prediction. Intelligent Control and Automation, 2024, 15: 9-27.

[9] Zhu, X., Huang, Y., Wang, X., Wang, R.: 'Emotion recognition based on brain-like multimodal hierarchical perception', Multimedia Tools and Applications, 2024, 83, (18), pp. 56039-56057.

[10] Tang Y., Wang Y., Zhou Q., et al. A New Approach to Risk Stratification for Heart Failure: Prognostic Value of Brain Natriuretic Peptide and Plasma Sodium Concentration. European Journal of Heart Failure, 2019, 21(6): 770-778.