Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/55/20241497

Heart disease exploration and prediction using data mining technique

Peien Ni

School of Future Science and Engineering, Soochow University, Suzhou, 215222, China

paytonenie@gmail.com

Abstract. In recent years, the incidence rate of various types of heart disease is increasing, so more and more studies focus on the influencing factors and prediction of heart disease. This paper investigates the relationship between heart disease and 21 representative variables to explore risk factors and predict the heart attack. First, through comparative analysis, we find that there is a significant difference in the distribution of variables. The findings highlight factors associated with heart disease, such as underlying diseases, lifestyle habits, and medical conditions. Secondly, these four factors (health condition, environment, age, and BMI) are extracted through factor analysis. Thirdly, a binary logistic regression prediction model is established to study the relationship between variables. Finally, the combination of multiple analysis methods in this study is a novel approach that can be used for exploring heart disease factors and predicting disease onset.

Keywords: Heart Disease, Data Mining, Comparative Analysis, Factor Analysis, Regression Analysis.

1. Introduction

The number of people who have suffered a heart attack has increased since 1961 [1]. Heart disease is one of the leading causes of death worldwide, especially in American [2], which not only affects the physical health of patients, but also brings a huge burden to society and the economy. For example, although the global incidence rate of congenital heart disease (CHD) is similar, countries with high fertility have a greater burden of heart disease to support patients, which aggravates the gap between countries [3]. In recent years, more researches focus on the prediction and diagnosis of heart disease in order to better prevent and treat the heart disease. Through these studies, there are many risk factors affecting heart disease. For instance, compared to men, diabetes, high-density lipoproteins, and triglycerides levels have been found to have a more significant impact on coronary heart disease risk in women [4]. Globally, High systolic blood pressure, high-density lipoprotein cholesterol, and smoking are the three main factors that are likely to cause heart disease [5]. Therefore, it is significant to explore the influencing factors of heart disease and predict its risk through data mining.

In 1973, Wilhelmsen et. al used a multiple logistic model to analyse nine probable risk factors and tested a logistic function in predicting coronary heart disease [6]. The techniques of meta-analysis to data were applied to exploring the impact of physical activity on coronary heart disease in 1990 [7]. Then, meta-regression models were fitted to evaluate non-linear effects of factors, for instance, alcohol

DOI: 10.54254/2755-2721/55/20241497

consumption [8]. In 2008, Decision Trees, Naive Bayes and Neural Network were used in a prototype Intelligent Heart Disease Prediction System conducted by Sellappan Palaniappan and Rafiah Awang [9]. As the meta-analysis continued to improve, an updated meta-analysis was created by Ramsden et. al to prevent the coronary heart disease in 2013 [10]. Kivimäki et. al (2015) used cumulative random-effects meta-analysis to combine effect estimates [11]. With the development of machine learning and deep learning, more new prediction models emerged in recent years. In 2019, the hybrid random forest with a linear model made the prediction have an accuracy level of 88.7% [12]. In the same year, a prediction model combining seven classification techniques occurred. Seven classification techniques are k-NN, Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machine, Neural Network and Vote. Its prediction accuracy reached 87.4% [13]. In 2022, supervised learning methods, SVM, Logistic regression, Decision Tree Classifier, Random Forest, and K- Nearest Neighbour were used in the analysis on an online UCI dataset to promote diagnosis and treatment [14]. In 2023, the effectiveness of deep learning in predicting heart disease has been proven to be superior to machine learning [15].

In this paper, we focus on 21 representative variables to study their relationship with heart disease through comparative analysis which is not frequently seen in previous studies. In addition, different from direct analysis of variables in some previous studies, we extract the main influencing factors through factor dimensionality reduction. Then a regression model of heart disease is established to predict whether heart disease will occur.

First of all, it is found that among the factors affecting heart disease, all discrete variables are independent of whether or not there is a disease, that is, the proportion of variables in the group with heart disease and the group without heart disease is significantly different. It is concluded that people with underlying diseases, poor lifestyle habits, and a lack of necessary medical conditions are more likely to suffer from heart disease. Studies on ordinal variables have shown that older, lower income, and lower education groups are more likely to suffer from heart disease. In addition, continuous variables fail the normality test, and it is found that they are independent of each other in the non-parametric test. This indicates that individuals with high BMI and poor mental and physical health are more likely to suffer from heart disease. Secondly, in the factor analysis, four factors are extracted by the maximum variance method, which are summarized as health condition, environment, age and BMI. Thirdly, the regression equation between 17 variables, such as hypertension and high cholesterol, and whether they suffer from heart disease has been established by binary logistic regression model. The accuracy of the regression model has been proved by calculating F1 scores, and it is found that Income and HvyAlcoholConsump are negatively correlated with heart disease attack. As with the previous analysis conclusion, the remaining variables are all positive feedback for heart attack.

The rest of the paper is organized as follows. Section 2 introduces the dataset and how to preprocess the dataset. In Section 3, we conduct the comparative analysis of a certain quantity between the two groups with or without heart disease, including the significance of the difference of discrete variables, the normality test of continuous variables, and the subsequent T-test and non-parametric test. Section 4 shows how to extract the main factors affecting heart disease through factor analysis. Section 5 sets up a binary logistic regression equation to predict whether a heart attack will occur and we conclude in Section 6.

2. Data Preprocessing

The Heart Disease Health Indicators Dataset, available on the Kaggle platform, is a large-scale dataset from the Behavioural Risk Factor Surveillance System (BRFSS). BRFSS is an annual health-related telephone survey conducted by the Centers for Disease Control and Prevention (CDC). This dataset includes 253,680 survey responses from cleaned dataset according to BRFSS 2015. There are 21 feature variables and a target variable with values of 0 and 1, representing the presence and absence of heart disease attack respectively. The explanations for variables are demonstrated in Table 1.

First, the data set is too large, although the dataset has been cleaned, we still checked some data records with missing variable values or duplicate data records. Then, divide the variables into scale,

DOI: 10.54254/2755-2721/55/20241497

nominal, and ordinal variables. Due to the dataset being encoded in advance, it is necessary to understand the data type and encoding method.

Regarding data coding, ordinal variables are Health General (GenHlth), Age, Education and Income, which respectively have 5, 13, 6, 8 grades. Then, scale variables are Body Mass Index (BMI), Mental Health (MentHlth), Physical Health (PhysHlth). Except for the above variables, variable Diabetes is ternary and the rest are binary nominal variables.

Table 1. Variable Explanations

Variables	Explanations	Variables	Explanations
HeartDiseaseorAttack	Binary	HvyAlcoholConsump	Men having more than 14 drinks weekly and adult women having more than 7 drinks weekly
HighBP	High Blood Pressure	AnyHealthcare	Any kind of health care coverage
HighChol	High Cholesterol	NoDocbcCost	Did you ever find yourself in a situation where you needed to consult a doctor within the last 12 months, but were unable to due to the cost?
CholCheck	Cholesterol examination within the past 5 years	GenHlth	The whole condition of health
BMI	Body Mass Index	MentHlth	Days of poor mental health
Smoker	The figure of cigarettes in whole life smoked is at least 100	PhysHlth	Days of poor physical health
Stroke	Binary	DiffWalk	Do you experience significant difficulty walking or climbing stairs?
Diabetes	Ternary	Sex	Binary
PhysActivity	Doing physical activity during the past 30 days other than their basis job	Age	Fourteen-level
Fruits	Eat fruit 1 or more times every day	Education	Six-level
Veggies	Eat vegetables 1 or more times every day	Income	Eight-level

3. Comparative Analysis

In order to explore the differences of different groups for a variable, descriptive statistical analysis and comparative analysis were used. Different test methods are selected for the significance of the differences between discrete and continuous variables.

3.1. Discrete Variables

3.1.1. Nominal Variables

For difference analysis of discrete variables, the chi-square test was chosen to test the significance of the difference. The chi-square test is a statistical test used to compare the difference between a metric in

DOI: 10.54254/2755-2721/55/20241497

two or more populations and can only be used for discrete variables. All nominal variables are analysed as discrete variables in this paper.

After the data is encoded, the dataset contains 14 nominal variables. As shown in Table 2, the difference in discrete variables between the two groups of customers (whether or not heart attack) was first examined. After the analysis of SPSS software, the ratio of each classification to whether have heart disease or not of the 14 variables and the proportion of each classification were obtained. At the same time, the chi-square test value and significance are also displayed, and it is used to determine whether the difference is significant.

Obviously, none of the 14 variables is independent of HeartDiseaseorAttack. Because the p-values of the chi-square tests are all less than 0.05, indicating the rejection of the null hypothesis, the difference between the heart-attack group and the non-heart-attack group in these 14 aspects is significant.

Specifically, based on the significant difference in terms of variables such as HighBP, HighChol, Stroke, Diabetes, DiffWalk, it is clear that the proportion of these variables with values of 1 or 2 is higher in people with heart disease. We can conclude that there is a significant relationship between the heart attack and underlying diseases.

Secondly, in terms of habits and customs, we can draw similar conclusions. The proportion of people who smoked at least 100 cigarettes in their entire life (Smoker) is about 42.5% in non-heart-attack group, smaller than that (61.9%) in heart-attack group. Adults who reported doing physical activity or exercise during the past 30 days other than their regular job in non-heart-attack group whose percentage is larger than that in heart-attack group, approximately 76.9%. In addition, the proportion of people who eat fruits 1 or more times every day in heart-attack group is smaller than that in non-heart-attack group. Similarly, the percentage of Fruits with a value of 1 heart-attack group is less than that in non-heart-attack group. Overall, people with good lifestyle habits have a lower proportion of heart disease.

Thirdly, with regard to medical treatment, we can find it that people who don't consult a doctor due to the cost are more likely to have heart disease, at about 11.1% (more than 8.1%).

Last but not least, the significant difference in gender ratio is reflected in the higher proportion of males in the population with heart disease (57.3%). Consequently, more care regarding heart disease prediction should be given to the male population. For instance, men have a higher incidence of heart failure [16].

Table 2. Results of 14 Pearson's Chi-square tests related to Heart Disease Attack.

Variables	**Classifications**	**Within HeartDiseaseorAttack**		**Total**	χ^2	**Asymptotic Significance(P)**
		0	**1**			
HighBP	0	60.4%	25.0%	57.1%	11117.883	.000
	1	39.6%	75.0%	42.9%		
HighChol	0	60.5%	29.9%	57.6%	8288.024	.000
	1	39.5%	70.1%	42.4%		
CholCheck	0	4.0%	1.1%	3.7%	494.932	.000
	1	96.0%	98.9%	96.3%		
Smoker	0	57.5%	38.1%	55.7%	3321.606	.000
	1	42.5%	61.9%	44.3%		
Stroke	0	97.2%	83.5%	95.9%	10450.577	.000
	1	2.8%	16.5%	4.1%		
Diabetes	0	86.3%	64.2%	84.2%	8244.889	.000
	1	1.7%	2.8%	1.8%		
	2	12.0%	33.0%	13.9%		
PhysActivity	0	23.1%	36.0%	24.3%	1932.628	.000
	1	76.9%	64.0%	75.7%		
Fruits	0	36.3%	39.5%	36.6%	99.215	.000

DOI: 10.54254/2755-2721/55/20241497

Table 2. (continued).

	1	63.7%	60.5%	63.4%		
Veggies	0	18.4%	23.6%	18.9%	388.824	.000
	1	81.6%	76.4%	81.1%		
HvyAlcoholConsump	0	94.2%	96.5%	94.4%	212.775	.000
	1	5.8%	3.5%	5.6%		
AnyHealthcare	0	5.0%	3.6%	4.9%	88.737	.000
	1	95.0%	96.4%	95.1%		
NoDocbcCost	0	91.9%	88.9%	91.6%	243.400	.000
	1	8.1%	11.1%	8.4%		
DiffWalk	0	85.7%	58.5%	83.2%	11475.802	.000
	1	14.3%	41.5%	16.8%		
Sex	0	57.3%	42.7%	56.0%	1879.793	.000
	1	42.7%	57.3%	44.0%		
Total	\	90.6%	9.4%	\	\	\

3.1.2. Ordinal Variables

There are four ordinal variables in the dataset: GenHlth, Age, Education, Income. Same as nominal variables, chi-square tests are also used. Calculated by SPSS, the p-values of the chi-square tests are all less than 0.05, which marks the rejection of the null hypothesis. As illustrated in Figure 1, in non-heart-attack group, self-perceived gentle health status concentrates at a high level. Similarly, we can know that education and income are also distributed in this way. It is clear that high income and education represent a higher level of medical care and awareness, promoting the prevention and treatment of heart disease. However, the population in non-heart-attack group is concentrated in the middle age while that in heart-attack group is concentrated in the elderly. This is because as one gets older, there are more and more underlying diseases.

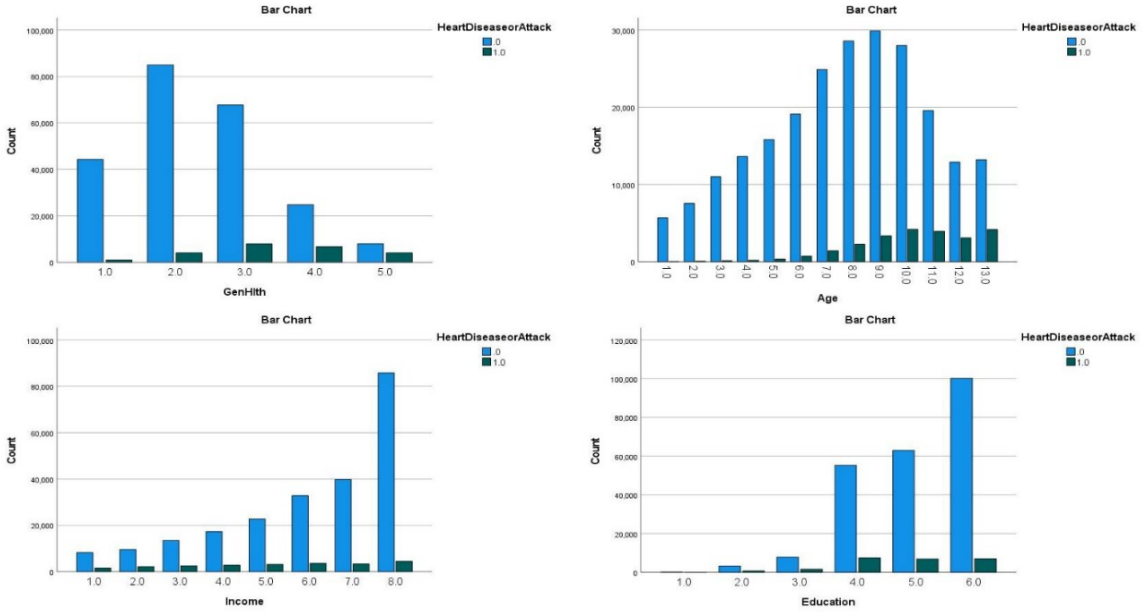

Figure 1. Bar Charts of Four Variables within Two groups.

DOI: 10.54254/2755-2721/55/20241497

3.2. Continuous Variables

Tests for difference significance of continuous variables require categorical discussion. If the distribution of the variables follows a normal distribution, you can use a t-test to compare differences in the mean. If the distribution of variables does not conform to a normal distribution, different nonparametric test methods should be chosen according to the number of groups X, and if X is two groups, such as sex, the Mann-Whitney U statistic should be used, and the Kruskal-Wallis statistic should be used if there are more than two groups.

In this data set, there are 3 continuous variables, which are BMI, MenHlth, PhysHlth. To begin with, the normality of all continuous variables is tested by the 'Explore' function in SPSS Software 'Descriptive Statistics'. As shown in Table 3, all continuous variables do not conform to a normal distribution. This is because all p-values are less than 0.05, which marks a rejection of the null hypothesis (the data conform to a normal distribution).

Naturally, nonparametric tests are chosen for these continuous variables. Fortunately, SPSS provides convenient nonparametric testing of independent samples. Similar to the chi-square tests, the nonparametric tests focus primarily on the difference in the mean of variables between heart-attack churn and non-heart-attack groups. As shown in Table 4, all decisions are rejecting the null hypothesis. It is clear that the distribution of the seven variables has respectively significant differences in the heart-attack churn and non-heart-attack groups.

Table 3. Results of Three Tests for the Normality of Continuous Variables.

Variables	**HeartDiseaseorAttack**	**Statistic**	**df**	**Sig.**
BMI	0	.120	229787	.000
	1	.109	23893	.000
MentHlth	0	.359	229787	.000
	1	.356	23893	.000
PhysHlth	0	.341	229787	.000
	1	.249	23893	.000

Table 4. Hypothesis Test Summary.

Null Hypothesis	**Test**	**Sig.**	**Decision**
The distribution of BMI is the same across categories of HeartDiseaseorAttack.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
The distribution of MentHlth is the same across categories of HeartDiseaseorAttack.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
The distribution of PhysHlth is the same across categories of HeartDiseaseorAttack.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

Thoroughly, it is necessary to explore the differences of each variable in detail through organized tables. As shown in Table 5 compared with the group that does not have a heart attack, the characteristics of the heart-attack group are obviously that BMI is higher and days of poor mental health is more as well as days of poor physical health. The numbers in parentheses indicate the magnitude of the change.

DOI: 10.54254/2755-2721/55/20241497

Table 5. Means and Std Errors of 3 variables within Two Groups.

Variables	Mean within HeartDiseaseorAttack		Std error	
	0	**1**	**0**	**1**
BMI	28.270	29.467	0.137	0.0436
MenlHlth	3.030	4.670	0.0150	0.0595
PhysHlth	3.731	9.154	0.0170	0.0768

4. Factor Analysis

Factor analysis aims to reduce a dataset's dimensionality by identifying hidden factors that explain the correlation structure among a set of variables. By uncovering the latent structure within the data, factor analysis can help researchers understand the fundamental relationships and patterns within a dataset. Based on the actual meaning of variables and the number of unique values, we can consider ordinal variables as continuous variables and add them to the model of factor analysis to improve the results of factor analysis.

Based on Table 6, the results of the KMO Test show that the value of KMO is 0.690. Meanwhile, the results of the Bartlett's Test illustrate that the P-value is much less than 0.01. This reaches the statistical significance, which indicates the rejection of the null hypothesis. In other words, the test shows that there is a correlation between the variables. Overall, Table 6 suggests that factor analysis can be conducted.

Table 6. KMO and Bartlett's Test.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**.690**
Bartlett's Test of Sphericity	Approx. Chi-Square	257025.307
	df	21
	Sig.	.000

Table 7 compares the initial communalities and extraction communalities of 7 variables respectively. The extraction communalities of all variables are greater than 0.6. Therefore, each of the original variables can be well explained by the extracted common factor.

Table 7. Communalities.

	Initial	**Extraction**
BMI	1.000	.957
GenHlth	1.000	.686
MentHlth	1.000	.687
PhysHlth	1.000	.713
Age	1.000	.884
Education	1.000	.778
Income	1.000	.694

Table 8 and Figure 2 demonstrates the total variance explained and the Scree Plot. As can be seen from the Scree plot, the eigenvalues of the common factors begin to decrease slowly from the component number 4. Therefore, four principal components should be extracted. Based on the Total Variance Explained, the first four common factors can explain 77.140% of the original data.

DOI: 10.54254/2755-2721/55/20241497

Table 8. Total variance explained.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.326	33.234	33.234	2.326	33.234	33.234	1.767	25.239	25.239
2	1.159	16.554	49.788	1.159	16.554	49.788	1.504	21.481	46.720
3	.979	13.984	63.772	.979	13.984	63.772	1.083	15.470	62.191
4	.936	13.368	77.140	.936	13.368	77.140	1.046	14.949	77.140
5	.629	8.987	86.127						
6	.531	7.585	93.711						
7	.440	6.289	100.000						

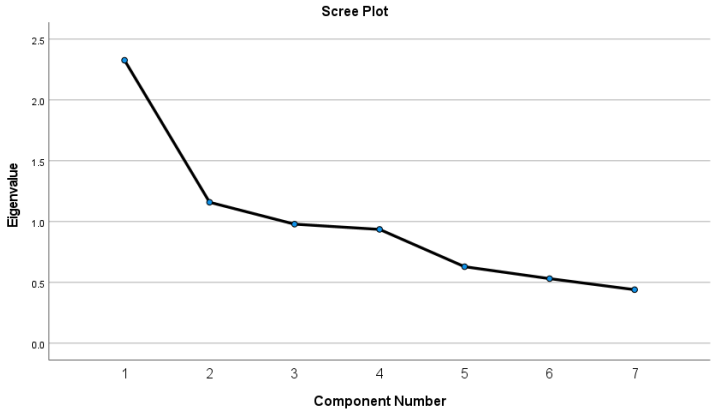

Figure 2. Plot.

Based on the Rotated Component Matrix (Table 9), this describes the factor loadings, i.e., the correlation coefficients of the variables with the common factors. After filtering the elements with factor loadings greater than 0.6, we determine the relevant variables that each common factor could represent well. Factor 1 can be named Health Condition, including GenHlth, MenHlth, PhysHlth. Factor 2 can be named Environment, including Education and Income. Factor 3 is Age and Factor 4 is BMI.

Table 9. Rotated Component Matrix.

	1	2	3	4
PhysHlth	.822			
MentHlth	.739			
GenHlth	.682			
Education		.879		
Income		.788		
Age			.934	
BMI				.972

5. Regression analysis

To predict whether a heart attack will occur, a binary logistic regression model between HeartDiseaseorAttack and other 21 variables is set up to study the relationship and to determine the degree and direction of influence of these independent variables on the dependent variable. Through

DOI: 10.54254/2755-2721/55/20241497

regression analysis, we can establish mathematical models, test the strength and direction of the relationship between independent and dependent variables, predict the values of the dependent variable, and identify the main factors affecting the dependent variable.

Through "Logistic Regression" in SPSS, we can get the classification table which is shown in Table 10. We can use the Accuracy, Precision, Recall and F1-score to evaluate the classification performance of the regression model. Through Table 10, we can know that TP(True Positive)is 18694, FP (False Positive) is 55745, TN (True Negative) is 174042, FN (False Negative) is 5199.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} = 76.0\% \quad (1)$$

$$Precision = \frac{TP}{TP+FP} = 25.1\% \quad (2)$$

$$Recall = \frac{TP}{TP+FN} = 78.2\% \quad (3)$$

$$F1 = \frac{2*Recall*Precision}{Recall+Precision} = 38.0\% \quad (4)$$

Based on the results of these four indicators, we can know that the classifier has a good classification effect.

Table 10. Classification Table.

Observed			HeartDiseaseorAttack		Percentage Correct
			0	1	
Step 16	HeartDiseaseorAttack	0	174042	55745	75.7
		1.0	5199	18694	78.2
	Overall Percentage				76.0

As shown in Table 11, 17 variables are selected into equation. The linear formula of logistic regression has been established in the equation (5). Except for the variable PhysActivity, the P-values of the Wald test for all other variables are less than 0.05, indicating that these regression variables have a significant impact on heart disease. The variable coefficients are mostly positive, indicating that the larger the variable value, the higher the likelihood of developing heart disease. However, the coefficients of Income are negative, indicating that the higher the income, the lower the likelihood of developing heart disease. For each level of income increase, the Odds Ratio (OR) value is 0.96 times that of the original. Significantly, both Stroke and Sex have significant advantages, with their Odds Ratio values being more than twice as high as before. This indicates that the stroke population has a higher probability of developing heart disease and that men are more likely to suffer from heart disease. Research shows that approximately 75% of stroke patients suffer from heart disease [17]. Therefore, in real diagnosis, doctors need to pay extra attention to stroke patients.

$$\ln\left(\frac{p}{1-p}\right) = -7.850 + 0.525 * HighBP + 0.612 * HighChol + 0.526 * CholCheck + 0.361 * Smoker + 0.979 * Stoke + 0.147 * Diabetes + 0.040 * PhysActivity + 0.047 * Veggies - 0.296 * HvyAlcoholConsump + 0.254 * NoDocbcCost + 0.494 * GenHlth + 0.003 * MentHlth + 0.303 * DiffWalk + 0.761 * Sex + 0.255 * Age - 0.041 * Income \quad (5)$$

Table 11. Variables in the Equation(Step 16).

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
HighBP	.525	.018	894.085	1	<.001	1.691
HlghChol	.612	.016	1383.876	1	<.001	1.843
CholCheck	.526	.066	63.293	1	<.001	1.691

DOI: 10.54254/2755-2721/55/20241497

Smoker	.361	.016	531.176	1	<.001	1.435
Stroke	.979	.024	1607.619	1	.000	2.661
Diabetes	.147	.009	280.386	1	<.001	1.159
PhysActMt/	.040	.017	5.508	1	.019	1.041
Veggies	.047	.018	6.430	1	.011	1.048
HeavyAlcoholConsump	-.296	.039	56.745	1	<.001	.744
NoDocbcCost	.254	.026	92.498	1	<.001	1.290
GenHlth	.494	.009	3233.630	1	.000	1.639
MentHlth	.003	.001	8.255	1	.004	1.003
DiffWalk	.303	.019	264.266	1	<.001	1.354
Sex	.761	.016	2268.118	1	.000	2.141
Age	.255	.003	5308.488	1	.000	1.290
Income	-.041	.004	108.369	1	<.001	.960
Constant	-7.850	.085	8601.459	1	.000	.000

6. Conclusion

Based on these experimental results, the main conclusion can be summarized as follows. Firstly, through comparative analysis, we analyse the differences in characteristic variables between populations with or without heart disease, and find that basic diseases are beneficial for clinical diagnosis of heart disease based on certain characteristics. Secondly, through factor analysis, we successfully divide continuous variables into four dimensions (health condition, environment, age, and BMI). In future research, these four dimensions require special attention. Finally, we establish a logistic regression model between all feature variables and HeartDiseaseorAttack, hoping to understand how feature variables affect the onset of heart disease. The most important limitation lies in the fact that the logistic regression prediction model has poor performance compared to some new machine learning models. We need to upgrade the prediction model to improve the accuracy of heart disease prediction.

Authors' Contributions

The author's contributions are fully considering the role of comparative analysis and factor analysis in exploring the influencing factors of heart disease and establishing a regression model to study the relationship between variables and heart attack.

Acknowledgments

First, I would like to express my sincere gratitude to David Woodruff from Carnegie Mellon University and Hudson Li from Peking University. They taught me algorithms for big data at the University of Chinese Academy of Sciences and shared their data mining expertise with me. I am deeply grateful for their assistance, which greatly benefited my academic pursuits.

My heartfelt thanks should also go to my parents for their continuous support and encouragement in terms of spirit and material.

Last but not least, I am indebted to my friends in the School of Future Science and Engineering. In particular, Yicheng Huang taught me a range of knowledge about machine learning whose earnest attitude triggered my pursuits of data science. In addition, my sincere thanks would go to my roommate Zaijun Fei for his company during difficult times.

DOI: 10.54254/2755-2721/55/20241497

References

[1] Scarborough, P., Wickramasinghe, K., Bhatnagar, P., & Rayner, M. (2011). Trends in coronary heart disease 1961–2011. London: British Heart Foundation, 2011.

[2] Fang, J., Luncheon, C., Ayala, C., Odom, E., & Loustalot, F. (2019). Awareness of heart attack symptoms and response among adults—United States, 2008, 2014, and 2017. Morbidity and Mortality Weekly Report, 68(5), 101.

[3] Hoffman, J. I. (2013). The global burden of congenital heart disease. Cardiovascular journal of Africa, 24(4), 141-145.

[4] Roeters van Lennep, J. E., Westerveld, H. T., Erkelens, D. W., & van der Wall, E. E. (2002). Risk factors for coronary heart disease: implications of gender. Cardiovascular research, 53(3), 538-549.

[5] Safiri, S., Karamzad, N., Singh, K., Carson-Chahhoud, K., Adams, C., Nejadghaderi, S. A., ... & Kolahi, A. A. (2022). Burden of ischemic heart disease and its attributable risk factors in 204 countries and territories, 1990–2019. European journal of preventive cardiology, 29(2), 420-431.

[6] Wilhelmsen, L., Wedel, H., & Tibblin, G. (1973). Multivariate analysis of risk factors for coronary heart disease. Circulation, 48(5), 950-958.

[7] Berlin, J. A., & Colditz, G. A. (1990). A meta-analysis of physical activity in the prevention of coronary heart disease. American journal of epidemiology, 132(4), 612-628.

[8] Corrao, G., Rubbiati, L., Bagnardi, V., Zambon, A., & Poikolainen, K. (2000). Alcohol and coronary heart disease: a meta-analysis. Addiction, 95(10), 1505-1523.

[9] Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In 2008 IEEE/ACS international conference on computer systems and applications (pp. 108-115). IEEE.

[10] Ramsden, C. E., Zamora, D., Leelarthaepin, B., Majchrzak-Hong, S. F., Faurot, K. R., Suchindran, C. M., ... & Hibbeln, J. R. (2013). Use of dietary linoleic acid for secondary prevention of coronary heart disease and death: evaluation of recovered data from the Sydney Diet Heart Study and updated meta-analysis. Bmj, 346, e8707.

[11] Kivimäki, M., Jokela, M., Nyberg, S. T., Singh-Manoux, A., Fransson, E. I., Alfredsson, L., ... & Virtanen, M. (2015). Long working hours and risk of coronary heart disease and stroke: a systematic review and meta-analysis of published and unpublished data for 603 838 individuals. The lancet, 386(10005), 1739-1746.

[12] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, 81542-81554.

[13] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, 82-93.

[14] Ramesh, T. R., Lilhore, U. K., Poongodi, M., Simaiya, S., Kaur, A., & Hamdi, M. (2022). Predictive analysis of heart diseases with machine learning approaches. Malaysian Journal of Computer Science, 132-148.

[15] Bakar, W. A. W. A., Josdi, N. L. N. B., Man, M. B., & Zuhairi, M. A. B. (2023, March). A Review: Heart Disease Prediction in Machine Learning & Deep Learning. In 2023 19th IEEE International Colloquium on Signal Processing & Its Applications (CSPA) (pp. 150-155). IEEE.

[16] Strömberg, A., & Mårtensson, J. (2003). Gender differences in patients with heart failure. European Journal of Cardiovascular Nursing, 2(1), 7-18.

[17] Barker, D. J. (1997). Intrauterine programming of coronary heart disease and stroke. Acta paediatrica, 86(S423), 178-182.