Research on Predictive Algorithms for Cardiovascular Disease

Yingzhen Wang*
Colorado State University, College Natural Sciences Department Statistic
wang1997@colostate.edu

ABSTRACT

Cardiovascular disease (CVD) is a global disease with acute and chronic complications. It is primarily responsible for the vast majority of deaths worldwide, which account for 17.9 million deaths annually. In terms of CVDs, illnesses like rheumatic heart disease and coronary heart disease are included, of which coronary heart disease (CHD) accounts for more than 50% of all these cases. In this research, principal component analysis (PCA) and backward stepwise elimination are used to identify the relevant predictors and avoid overfitting models for random forest analysis and logistic regression analysis. Moreover, for assessing the effectiveness of the models, the confusion matrix and the receiver operating characteristics (ROC) curve with AUC (area under the ROC curve) value are produced for model comparison. The outcomes demonstrate that the random forest model performs better at categorizing high-dimensional data. Thus, the techniques discussed in this paper give medical researchers better ways to handle coronary heart disease data statistically and provide a new statistical procedure for coronary heart disease prediction and prevention.

CCS CONCEPTS

• **Probability and statistics**; • **Machine learning**; • **Life and medical sciences**;

KEYWORDS

Coronary heart disease, Principal component analysis, Random Forest, Logistic Regression

ACM Reference Format:

Yingzhen Wang*. 2023. Research on Predictive Algorithms for Cardiovascular Disease. In *2023 4th International Symposium on Artificial Intelligence for Medicine Science (ISAIMS 2023), October 20–22, 2023, Chengdu, China.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3644116.3644169

1 INTRODUCTION

Cardiovascular diseases (CVDs) are widespread worldwide and more than 485.6 million patients had CVDs in 2017 [1]. CVD has a high fatality rate according to worldwide investigation. In terms of mortality in the world's population, CVDs were responsible for 31.8% of deaths globally in 2017 [2]. CVD is accomplished with multiple acute complications, including acute myocardial infarction, heart attack and the most common one, coronary heart disease (CHD). To improve bodily signs, sustaining or switching to healthy lifestyles can significantly lessen the fatality of these diseases. In a study on second-degree preventive measures, advocating a Mediterranean-inspired dietary habit for four years reduced the risk of sudden cardiac death by 72% [3]. Each standard deviation decrease in low-density lipoprotein cholesterol levels was associated with a 21% reduction in the incidence of an acute myocardial infarction [2]. The pathogenicity of CHD is linked to numerous intricate factors. Targets for lowering the likelihood of heart disease have been recognized to be the two primary risk factors for CHDs, which are hypertension and diabetes [3]. However, some unnoticed lifestyles are also the major contributing factors to CHD. In other words, lifestyle has a substantial influence on risk factors like excessive blood sugar, elevated body mass index, and high degrees of low-density cholesterol lipoprotein. [4].

Tobacco products are highly industrialized and contain the highly addictive nicotine that includes toxic substances and carcinogens [5]. Smoke from cigarettes contains chemicals that can thicken blood and lead to clots in arteries and veins, which accelerates plaque formation in blood vessels. When the arteries that provide blood to the heart muscle are constricted by plaque or obstructed by a clot, coronary heart disease develops. There were 17.5 million deaths from CVDs in 2012 where smoking is a particular cause [6]. More over 34 million adults, or 14 percent of the public, still smoking in 2017, despite the fact that the consumption of tobacco is declining in the western world [7]. Along with smoking, obesity is also a major contributor to CHD. Since the contemporary high-stress population is frequently physically inactive and unconsciously prefers to eat high-sugar and high-calorie edible products, the number of CHD cases due to obesity is increasing per year. Obesity results in abnormally high amounts of cholesterol and glucose in blood and organs and insulin resistance, which directly influences the occurrence of coronary risk factors such as diabetes and hypertension. Cardiovascular death is 50–60% more likely to occur in diabetic patients than in non-diabetic patients [8, 9]. High blood pressure causes myocardial functional and structural disturbances and increases cardiac workload [10]. One of these alterations that might result in cardiac failure is the muscular expansion of the left atrium. One of the most crucial metrics for determining the likelihood of developing CHD is the body mass index (BMI). Obesity, which increases the risk of CHD, is defined by a BMI of 30 kg/m2 or higher. According to a substantial amount of research evidence, the most effective strategy for CHD prevention and control is maintaining a healthy lifestyle. These lifestyles include, but are not limited to, maintaining a healthy weight through calorie restriction, having a proper diet, abstaining from alcohol and tobacco, and increasing daily physical activity [4]. Alterations to lifestyle may be the most crucial aspect of primary CHD prevention and management, in consideration of the advantages to the

ISAIMS 2023, October 20–22, 2023, Chengdu, China

ACM ISBN 979-8-4007-0813-8/23/10
https://doi.org/10.1145/3644116.3644169

Table 1: Variable description

Variable	Description	Type 0	Type 1
male	gender of the subjects	female	male
currentSmoker	whether the subject is currently smoking or not	no	yes
BPMeds	whether the subject took any blood pressure medicine	no	yes
prevalentStroke	whether the subject had ever undergone a stroke	no	yes
prevalentHyp	if the subject had hypertension or not	no	yes
diabetes	if the subject had diabetes or not	no	yes
CigsPerDay	the average daily cigarette consumption for the subject	no	yes
totChol	total cholesterol measurement	no	yes
sysBP	systolic blood pressure	no	yes
diaBP	diastolic blood pressure	no	yes
BMI	body mass index	no	yes

population and the reduction of harmful pharmacological effects [3].

This article utilizes logistic regression and random forest to statistically analyze more than 4,000 citizens of the Massachusetts region of Framingham, determining whether a patient has a 10-year likelihood of developing coronary heart disease utilizing the subject's biological markers and clinical conditions as predictors. The objectives of this paper are to figure out the most effective predictive statistical model and calculate the total risk (odds) of developing CHD.

2 METHODS

2.1 Data Sources and Description

4,239 observations and 16 variables from a cardiovascular study that is currently being conducted on citizens of the state of Massachusetts city of Framingham serve as the main source of the original data. Each of the 15 independent variables can be treated as clinical, physiological, and sociological predictors. The "TenYearCHD" prediction target is a binary variable with two classes: 0 indicates there is no decade chance of developing coronary heart disease, and 1 implies there is a ten-year possibility of coronary heart disease for the subject.

2.2 Variable Information

In the follow-up research and model fitting, some variables are encoded. Predictor "Education" has been changed into factors by each corresponding level, including 1 (not completing high school), 2 (completing high school), 3 (completing college), and 4 (completing post-college). The rest of the encoding variables and several shortened continuous variables are as Table 1 shows:

2.3 Research Methods

The main model fitting methods used in the following research include logistic regression and random forest. Following model fitting, the effectiveness of two models is evaluated using the matrix of confusion and ROC curve. Among them, random forest and logistic regression are two effective classification algorithms to handle analysis problem with binary dependent variable, predict the odds of having coronary heart disease (CHD) by each predictor, and identify the factors that can be mitigated and prevented by improving lifestyle.

2.3.1 Random Forest. A widely used consensus algorithm called random forest comprises multiple decision trees for both regression and classification problems. The ensemble of decision trees is trained through the Classification and Regression Tree (CART) algorithm to produce a single prediction result. To produce a more precise estimate, random forest employs an ensemble technique known as bagging.

Feature randomness implemented in random forest ensures low correlation among decision trees, which reduces the risk of bias, and the ensemble of decision trees avoids the problem of overfitting. These two techniques enable random forest to process high-dimensional data and guarantee the high accuracy of predictions.

2.3.2 Logistic Regression. Frequently used for classification and forecasting techniques is the logistic regression model. The likelihood of an event occurring is calculated using logistic regression using a number of predictors. In logistic regression, the odds are transformed using the logit function.

A common approach to estimating the coefficients in a logit model is maximum likelihood estimation (MLE). Through several iterations, to determine the coefficients that best fit the log odds, MLE assesses several coefficient values. Following the identification of the optimal coefficients, it is possible to log and sum the conditional probabilities for each data point to obtain the predicted odds.

2.3.3 Confusion Matrix. A matrix for summarizing the performance of classification algorithm is called confusion matrix. It has two rows and two columns including the counts from predicted and actual values. For instance, false positives represent the amount of real negative observations that were mistakenly categorized as positives, while true negatives represent the quantity of actual negative observations that were correctly classified as negative.

Several indicators for assessing model performance can be calculated using just the four parameters from the confusion matrix. Accuracy is the proportion of the correct predictions from the model. The percentage of accurately identified real positive cases is known as sensitivity and refers to the model's ability to designate an observation with "event = 1" as positive. Specificity is the proportion of

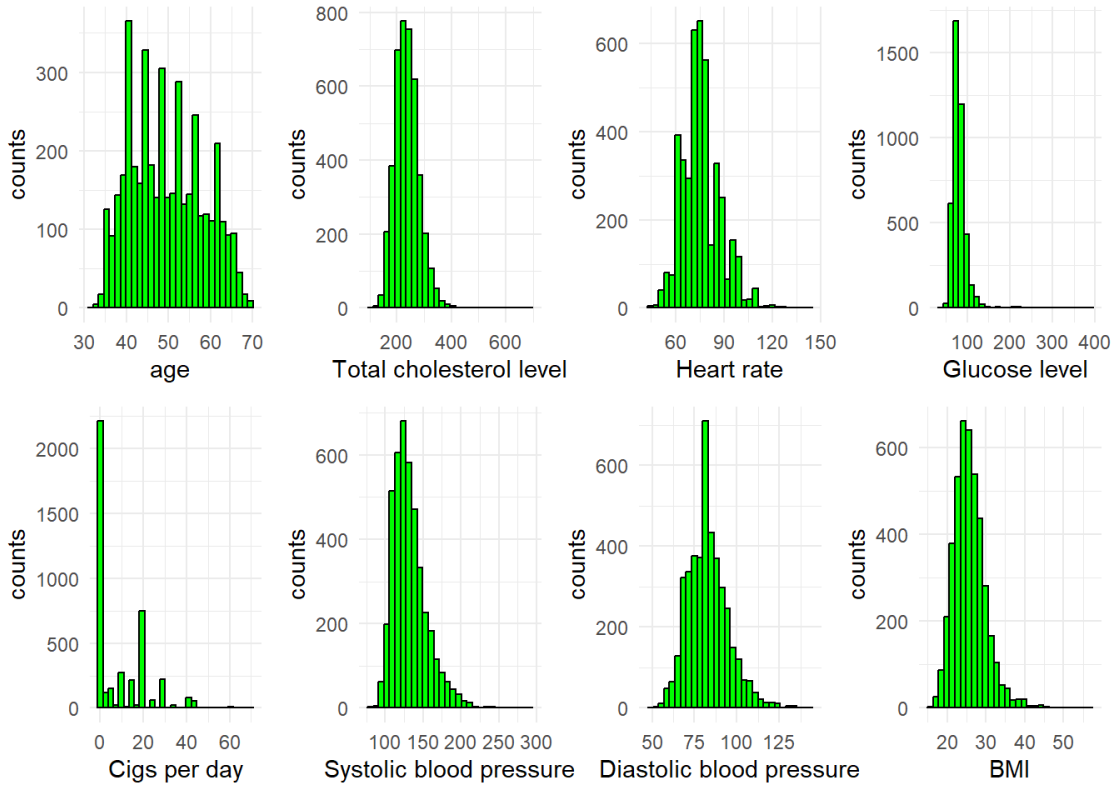

Figure 1: Density plots of continuous variables

correctly classified actual negative cases and refers to the model's ability to classify an observation with "event = 0" as negative.

2.3.4 AUC-ROC Curve. AUC - ROC curve is one of the most effective evaluation metrics for measuring the performance of a classification model at various threshold values. A probability curve called the ROC is plotted between the sensitivity and 1 minus specificity. AUC is used to calculate an aggregate measure of separability that represents the performance of the binary classification algorithm across several thresholds. AUC of a model near to 1 means the model has a great measure of separability and hence excellent performance, while AUC equals to 0.5 indicates a model has no classification ability.

3 RESULTS AND DISCUSSION

Coronary heart disease is a serious chronic condition that weakens the blood supply capacity of a person's heart muscle, triggers a range of clinical diseases, and shorten a person's quality of life and life expectations [1]. When under physical activity or mental stress, a person's heart needs more oxygen-rich blood [8]. In response to this signal, the blood arteries enlarge, allowing for increased blood flow. However, a person's unhealthy lifestyles and other conditions (stroke or diabetes) cause atherosclerosis and hence narrow the blood vessels.

3.1 Exploratory Analysis

In the process of integrating data from the ongoing cardiovascular study, certain rows ended up missing a total of 645 data points for various reasons. Since these rows contain the original data in addition to the missing values, simply removing these rows will result in unreasonable final analysis results. Therefore, Multivariate Imputations by Chained Equations (MICE) is applied to deal with this issue, because it can create multivariate imputations and impute mixes of continuous and categorical data.

The normality of continuous variables is of great significance for statistical research. As shown in Figure 1, all the continuous variables represent an approximate normal distribution.

Another critical factor in statistical research is outliers. Based on the boxplots shown in Figure 2, there are plenty of outliers within several continuous variables. These outliers affect the mean values of the dataset and the precision of the analysis.

To handle the outliers, the capping algorithm is an effective and prevalent tool. Specifically, the value of the 5% quantile is substituted for observations that fall outside the lower limit, while the value of the 95% quantile is substituted for data that fall outside the upper limit [11]. After implementing the algorithm to each variable with outliers, such problem is settled as shown in Figure 3.

Among all the subjects, those with ten-year risk of coronary heart disease are accounted for 3594 (84.80%) and others without this risk are accounted for 644 (15.20%), which indicates that the

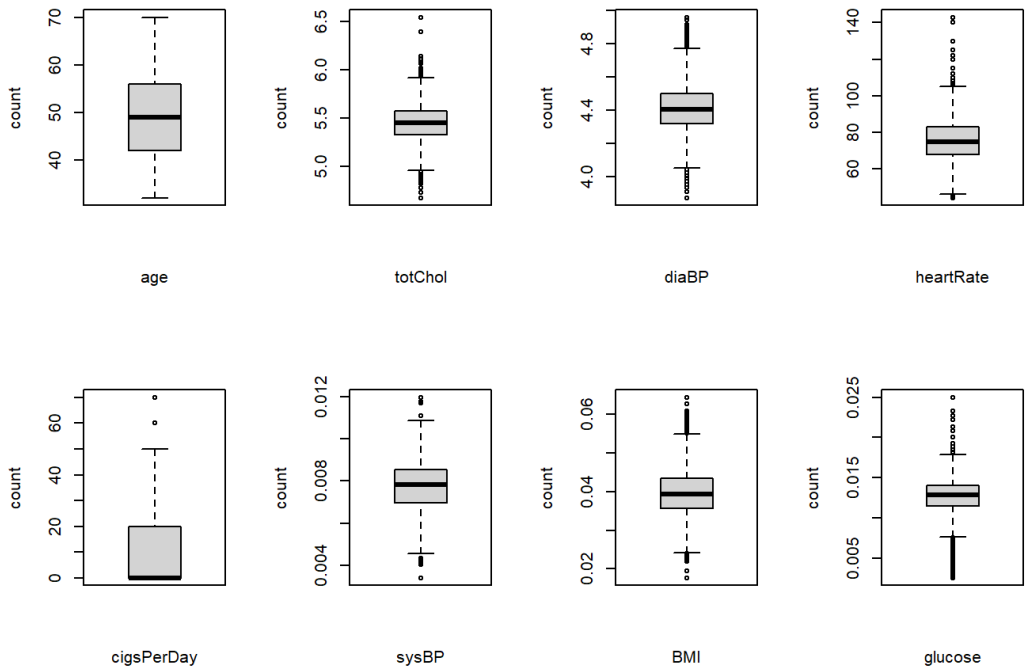

Figure 2: Boxplots of continuous variables

Table 2: Distribution of the remaining categorical variables by classes

Variable by Classes	TenYearCHD = 0	TenYearCHD = 1	Counts
TenYearCHD = 0	2123	NA	2123
TenYearCHD = 1	NA	2115	2115
male = 0	1281	977	2258
male = 1	842	1138	1980
currentSmoker = 0	1105	1027	2132
currentSmoker = 1	1018	1088	2106
BPMeds = 0	2058	1977	4035
BPMeds = 1	65	138	203
prevalentStroke = 0	2112	2079	4191
prevalentStroke = 1	11	36	47
prevalentHyp = 0	1544	1070	2614
prevalentHyp = 1	579	1045	1624
diabetes = 0	2085	1974	4059
diabetes = 1	38	141	179

dataset is unbalanced. Due to the binary classification job involved in this study, it is suitable to create synthetic data for the raw data by randomly over sampling examples (ROSE) [12].

Education is a categorical variable with 4 levels, and the proportions of "TenYearCHD" in each level are demonstrated in a bar plot. As shown in Figure 4, approximately 56.63% subjects in less than high school education background group have the ten-year CHD risk.

The proportion of the remaining categorical variables in the data set corresponding to their classes versus the prevalence by "TenYearCHD" is shown in Table 2. Based on this table, proportions of each class can be calculated according to the target variable's classes.

As shown in Table 2, eight continuous predictors, seven categorical predictors, and one target binary variable are included in the data set, and it is balanced after over sampling. Before model fitting,

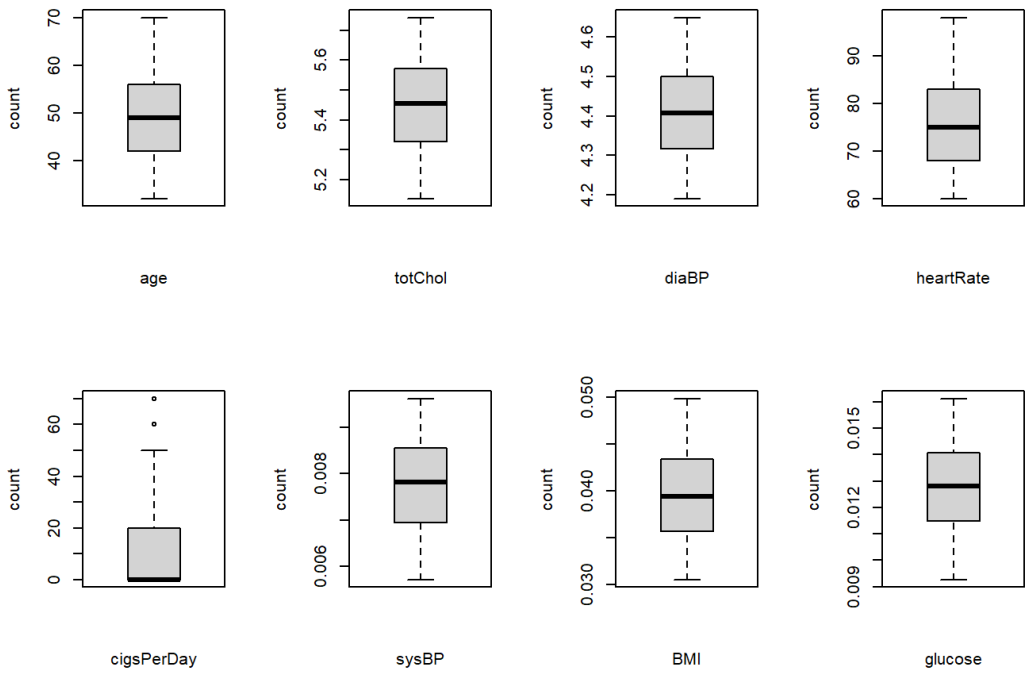

Figure 3: Boxplots of continuous variables after capping

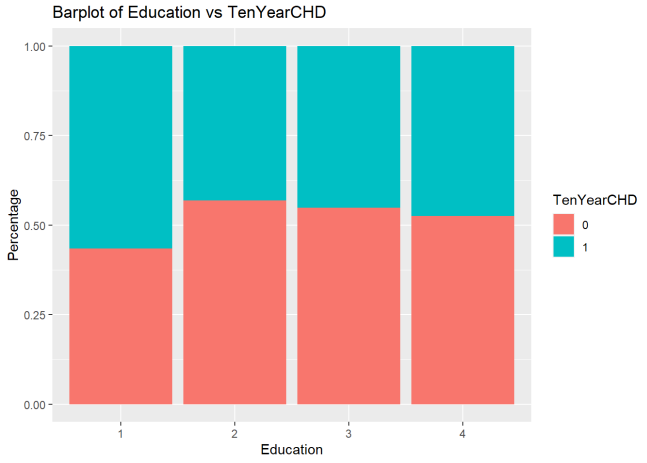

Figure 4: Bar plot of variable education

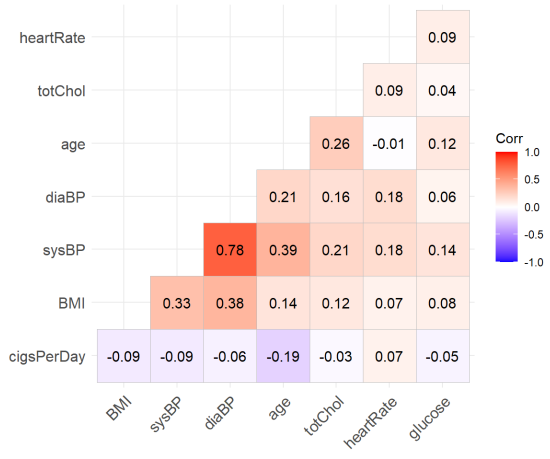

Figure 5: Correlation plot among continuous variables

it is necessary to check the collinearity between each continuous variable.

As shown in Figure 5, there is a strong positive correlation (0.78) between sysBP and diaBP. In order to determine whether these two predictors are collinear, a scatter plot with a fitted line is created.

As shown in Figure 6, sysBP and diaBP are significantly positively correlated with one another, which indicates collinearity between these two variables. Therefore, these two predictors should not appear in the same model simultaneously as possible.

3.2 Random Forest Results

In statistical modeling, variable selection is a significant step before modeling. After selecting the most critical variables and eliminating the redundant ones, the prediction power of the model can be improved.

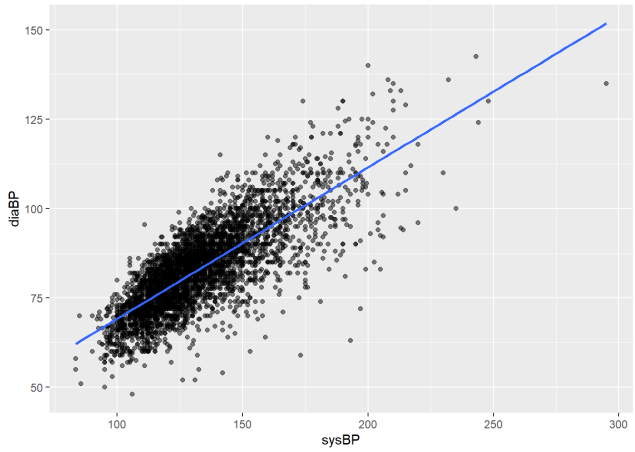

Figure 6: Scatter Plot for sysBP and diaBP

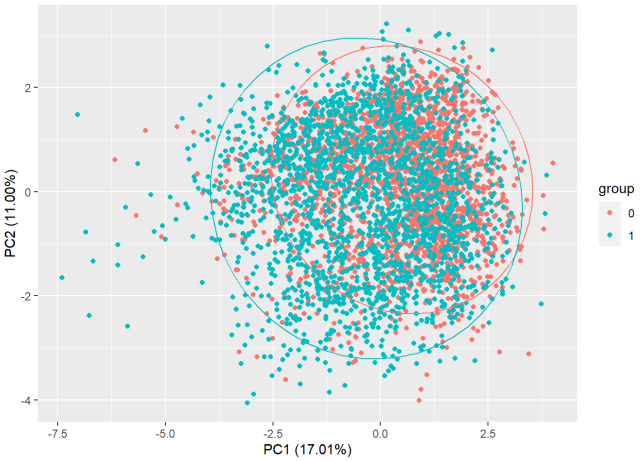

Figure 8: Principal component plot

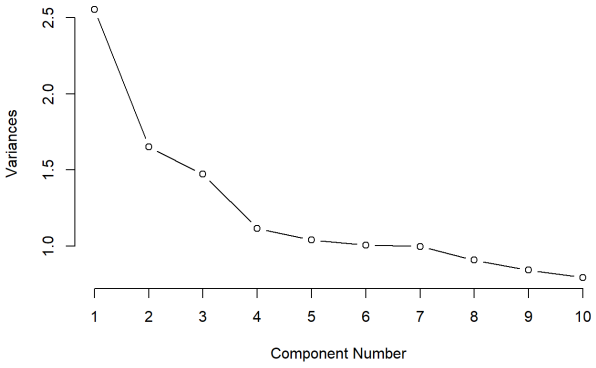

Figure 7: Gravel plot

3.2.1 Principal Components Analysis. Before fitting the random forest model, principal components analysis (PCA) is performed on 15 variables, and selected the predictors related to the target variable. The noise of the data set can be reduced by PCA through variable selection and dimension reduction based on the target variable [13]. In order to select principal components with high degree of interpretation, a gravel plot is drawn according to variance of each eigenvalue and contribution rate of each principal component to variance.

As elucidated in Figure 7, the relationship between eigenvalues and the number of principal components is evaluated using the Cattell gravel test [13]. The eigenvalue of component 6 is great than 1 and the eigenvalue of parallel analysis. It is worth noting that the eigenvalue of component 2 is much higher than the rest 4 components. Therefore, choosing the first two principal components is optimal due to high degree of interpretation of the data set.

As illustration in Figure 8, PC1 (17.01%) and PC2 (11.00%) explains approximately 28.01% of the variance of 15 variables, which means that most of the information can be interpreted by using these two components. For variable selection, the criteria are to select the variables with the relatively high absolute weights in PC1 and PC2 calculations. The weights indicate the degree of contribution to the linear combination. Based on this criteria, age (0.29), prevalentHyp (0.45), prevalentStroke (0.47), sysBP (0.49), diaBP (0.43), totChol (0.20), diabetes (0.23) and BMI (0.28) are selected based on their weights in PC1 calculation, and male (0.37), heartrate (0.23), currentSmoker (0.57), and cigsPerDay (0.61) are selected based on their weights in PC2 calculation. However, there is a strong collinearity between sysBP and diaBP, hence sysBP is selected based on its higher weight, and cigsPerDay and currentSmoker are correlated with each other, hence cigsPerDay is selected based on its higher weight. In conclusion, age, prevalentHyp, prevalentStroke, sysBP, totChol (total cholesterol level), BMI, diabetes, heartrate, male, and cigsPerDay are the predictors for random forest model fitting.

3.2.2 Random Forest Model Fitting and Analysis. In the process of model optimization, a grid search algorithm is implemented to adjust the depth and number of the trees. In this algorithm, the ntree parameters of the random forest model are 100-500 step as 100, and max.depth parameters are 2-20 step as 2. The training set contains 70% of the samples, and the validation set contains 30% of the samples. Following multiple iterations to train the model, the random forest model with ntree = 200 and max.depth = 6 has the highest prediction accuracy.

In the model fitting part, the random forest model fitted with the target variable and the predictors selected from PCA analysis. These predictors include previous health status, bio status indicators, and other basic information of the subjects. Interpreting the importance ranking of the predictors is necessary for understanding the relationship between predictors and target variable.

As displayed in Figure 9, the importance of predictors is determined using the MeanDecreaseGini technique [13]. Based on these results, sysBP, BMI, and totChol are the top three predictors with the strong association with ten-year future risk of coronary heart disease. Age, heartrate, and cigsPerDay with moderate importance

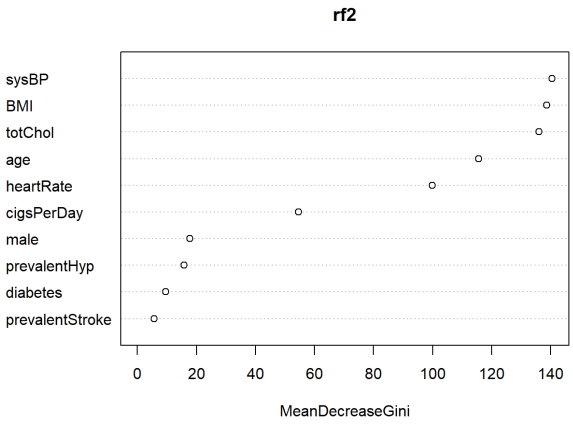

Figure 9: Importance plot

are also regarded as the risk factors for ten-year future risk of coronary heart disease. Predictors such as male, prevalentHyp, diabetes, and prevalentStroke are at relatively lower ranks, probably due to their relatively weak relationship with ten-year likelihood of CHD. In fact, ratings for the predictors are only the importance derived from random forest model. Identifying key risk factors needs to be combined with practical scenarios.

3.3 Logistic Regression Results

Likewise, model selection is also employed to determine the optimal logistic regression model. Nevertheless, the statistical significance of estimates in logistics regression is important, hence the backward stepwise elimination is used to perform model selection [10].

3.3.1 Backward Stepwise Elimination. In backward stepwise elimination, the predictors in saturated model reduce based on a process that gradually removes the least significant (highest p-value) variable in each step. The process is repeated until the remaining predictors are significant [10]. This approach can effectively avoid the overfitting issue and improve the predicting power. In this research, the significance level is set as the default value 0.05 and using 15-predictor full logistic regression model to perform the backward stepwise elimination. In order to ensure the results, the corresponding AIC and BIC values are calculated as the reference.

As shown in Table 3, the quality of each logit model is assessed using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). In general, lower AIC and BIC indicates a better model. The optimal logistic model with a binomial logit link contains predictors such as male, age, cigsperday, prevalentHyp, diabetes, totChol, and sysBP. Although the AIC value of this model is slightly higher than the previous one, the BIC value is the smallest one and all these predictors are significant (p-value < 0.05). It should be noted that the selected significant predictors are only the outcomes based on statistical method for statistical analysis. Combining complicated practical conditions is needed to make the selection more comprehensive and accurate.

3.3.2 Diagnostic Analysis and Estimates Interpretation. The nonexistence of multicollinearity, the linear relationship between logit and continuous variables, and the lack of outliers with a high level of influence are the top three assumptions for logit models [9]. To check these assumptions, three diagnostic methods are implemented, including logit versus predictor values plot, cook's distance plot, and VIF test.

As presented in Figure 10, values of age, cigsPerDay, sysBP, and totChol have positive linear relationships with logit, which indicates that the assumption of linearity in the logit for continuous variables is reasonable.

According to the cook's distance shown in Figure 11, number 653, 1392, and 1784 are three influential observations. However, influential observations are not necessarily influential outliers. Based on the criterion that observations with absolute standardized residuals above 3 should be regarded as outliers, the three influential observations are not influential outliers. Thus, the assumption of lack of strongly influential outliers is tenable.

The Variance Inflation Factor (VIF) assesses the degree of correlation among the features. Specifically, the VIF score of a predictor exceeding 10 is a sign of severe multicollinearity among this predictor and the remaining predictors. According to the results of VIF test to the optimal logistic model, all predictors have VIF scores around 1, which indicates absence of multicollinearity.

As shown in the Table 4, all the parameter estimates are transformed into exponential form for more practical interpretation. For the exponential estimate of prevalentHyp1, the odds of previous hypertensive subjects developing coronary heart disease within

Table 3: Backward stepwise elimination results

Predictor	AIC	BIC
all 15 predictors	5213	5328
without diaBP (p-value = 0.96)	5211	5319
without diaBP and BPMeds (p-value = 0.89)	5209	5311
without diaBP, BPMeds, and glucose (p-value = 0.84)	5207	5303
without diaBP, BPMeds, glucose, and heartrate (p-value = 0.62)	5206	5294
without diaBP, BPMeds, heartrate, glucose, and currentSmoker (p-value = 0.47)	5204	5287
without diaBP, BPMeds, heartrate, glucose, currentSmoker, and prevalentStroke (p-value = 0.18)	5203	5280
without diaBP, BPMeds, heartrate, glucose, currentSmoker, prevalentStroke, and BMI (p-value = 0.11)	5202	5275
without diaBP, BPMeds, heartrate, glucose, currentSmoker, prevalentStroke, BMI, and education (p-value = 0.27)	5203	5256

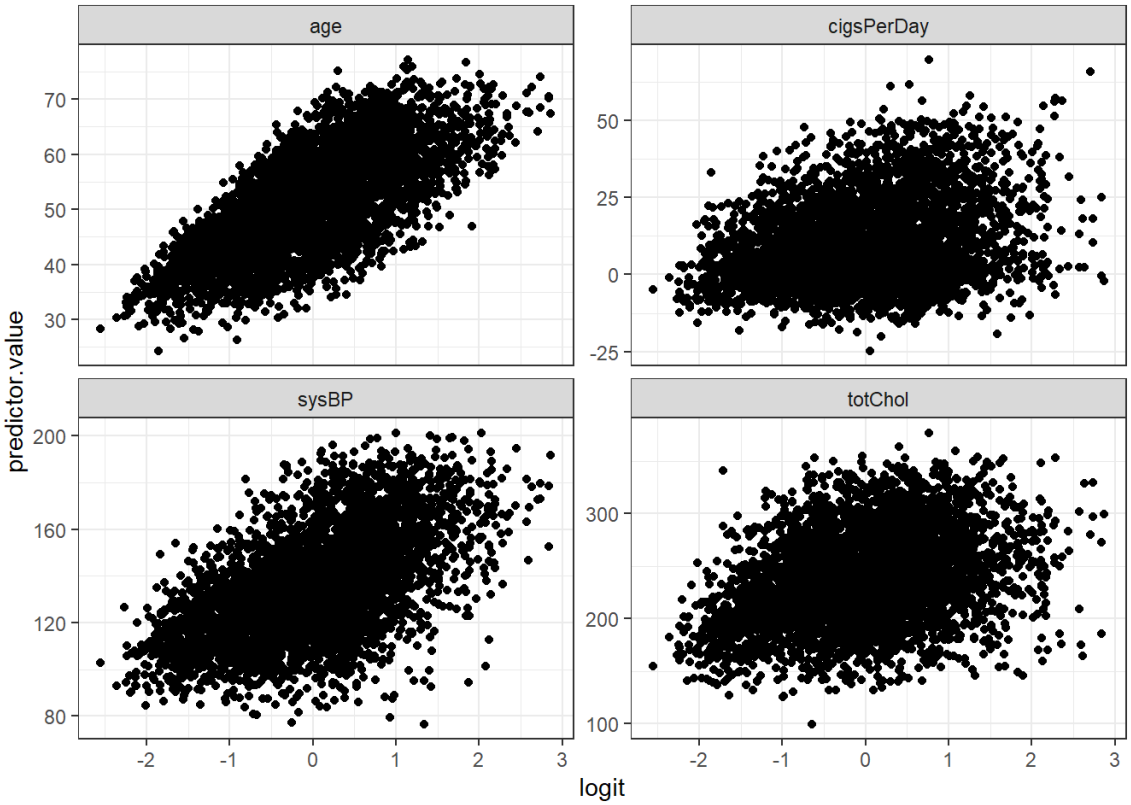

Figure 10: Logit vs. predictor values plot

Table 4: Coefficient estimates

Predictor	Estimate	Exponential Estimate	P-value
intercept	-5.333	0.005	$< 2 \times 10^{-16}$
male1	0.563	1.756	1.38×10^{-15}
age	0.054	1.056	$< 2 \times 10^{-16}$
cigsPerDay	0.019	1.019	1.05×10^{-12}
prevalentHyp1	0.475	1.609	1.60×10^{-7}
diabetes1	1.026	2.791	1.07×10^{-7}
totChol	0.003	1.003	1.08×10^{-5}
sysBP	0.008	1.008	7.47×10^{-5}

the next decade is approximately 1.609 times higher than that of subjects without hypertension. Moreover, the odds of having coronary heart disease in the future ten years increase by approximately 1.90% when smoking an extra cigarette per day. The exponential estimates of other predictors can be interpreted as the same patterns as above.

3.4 Model Comparison Results

To determine the model with higher forecasting accuracy and better classification capability, the confusion matrix and ROC curve with AUC value are employed in this section. Several indicators derived from these two methods have the function of evaluated model performance.

As shown in Figure 12, the accuracy can be calculated to be 0.671, the sensitivity is 0.664, and the specificity is 0.680. Among them, in the optimal logistic model, 67.1% predictions are correct, 66.4% actual positive cases are correctly classified, and 68% actual false cases are correctly classified.

As shown in Figure 13, the AUC value is 0.737. Based on all these indicators, the overall performance of the optimal logistic model is relatively not outstanding, and its ability to classify binary target variable is at a moderate level. Therefore, evaluating the optimal

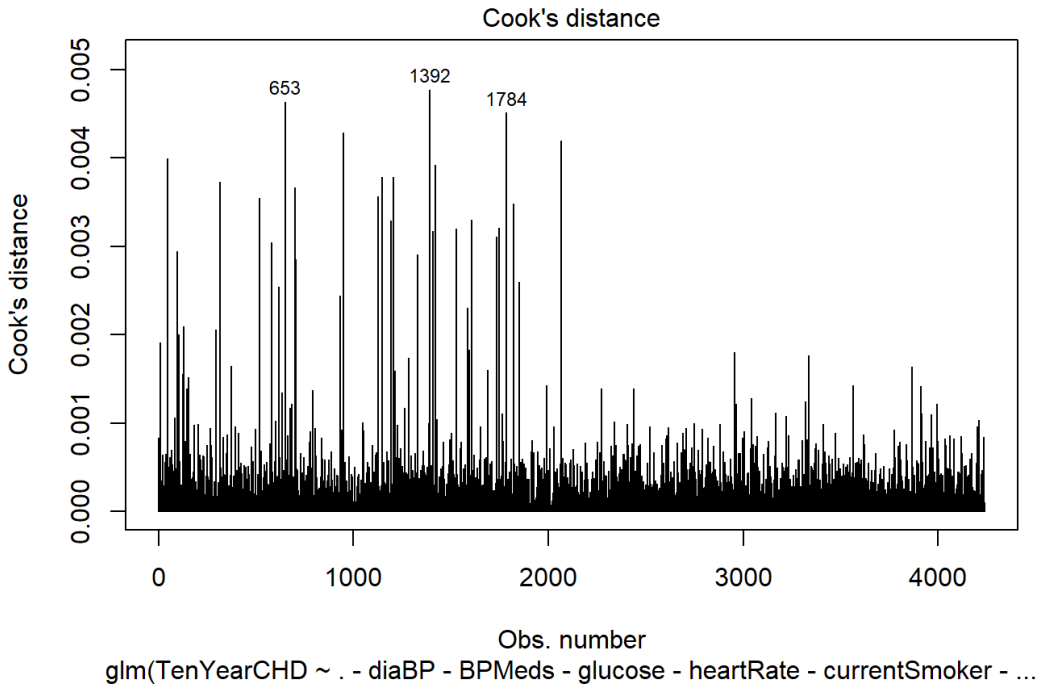

Figure 11: Cook's distance plot

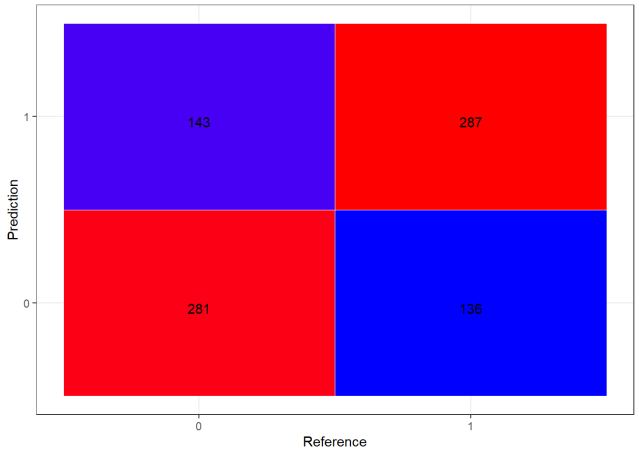

Figure 12: Confusion matrix for the optimal logistic model

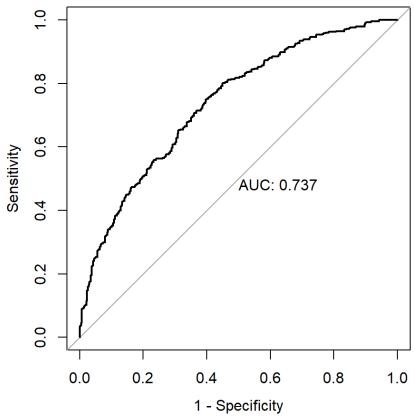

Figure 13: AUC-ROC curve for the optimal logistic model

random forest model performance may yield a more satisfactory result.

Based on the information of Figure 14, in the optimal random forest model, 82.1% predictions are correct, 83.1% actual positive cases are correctly classified, and 80.9% actual false cases are correctly classified. As compared to the optimal logistic model, the accuracy is 15% higher, the sensitivity is 16.7% higher, and the specificity is 12.9% higher.

As shown in Figure 15, the optimal random forest model's AUC is 0.82 and is considerably closer to 1, which indicates that this model has more classification ability than the optimal logistic model has. In conclusion, the optimal random forest model provides more accurate predictions and has greater categorization ability. The models used in this section are processed by feature selection and

Table 5: Classifier Performance

Classifier	Accuracy	Sensitivity	Specificity
K-Nearest Neighbors (KNN)	0.75	0.83	0.68
Gradient Boosting (GB)	0.73	0.76	0.70
Support Vector Machine (SVM)	0.67	0.71	0.64
Naive Bayes	0.61	0.36	0.85
XGBoost	0.70	0.70	0.69
Multilayer Perceptron (MLP)	0.67	0.67	0.68
Logistic Regression (In this research)	0.67	0.66	0.68
Random Forest (In this research)	0.82	0.83	0.81

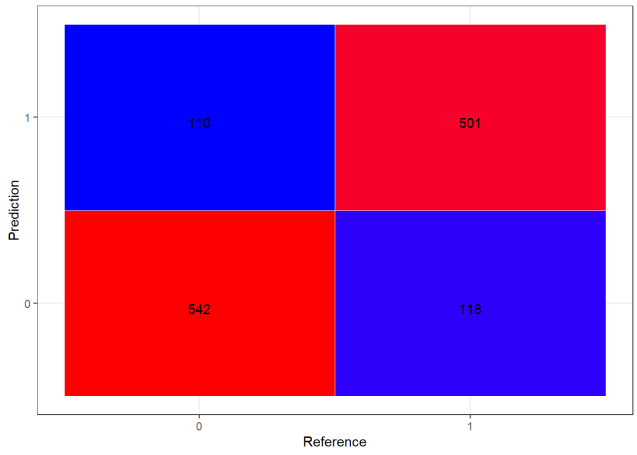

Figure 14: Confusion matrix for the optimal random forest model

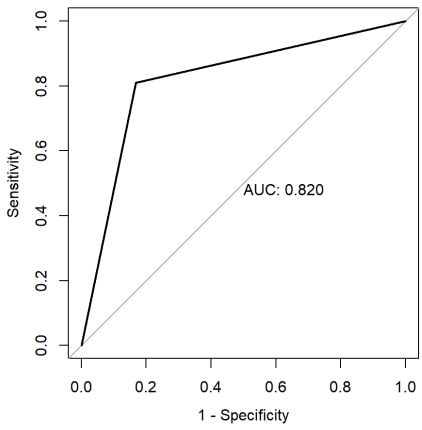

Figure 15: AUC-ROC curve for the optimal random forest model

model selection, hence the results are reliable and scientific to some extent.

For the comprehensiveness of the research, other types of predictive algorithm are analyzed as reference for model comparison. Particularly, six other algorithms are implemented as classifiers into this dataset for statistical predictions [14, 15]. By comparing these classifiers with the two predictive algorithms in this study, the conclusions drawn will be relatively accurate and trustworthy. According to the accuracy, sensitivity, and specificity of these algorithms, a table is created to compare the performance of the classifiers [14, 15].

As illustrated in the Table 5, random forest classifier has the highest accuracy, sensitivity, and specificity, which indicated that this classifier is more suitable for analyzing this data set and has the ability to guarantee the accuracy of classification.

4 CONCLUSION

Throughout this research, the risk factors of ten-year future risk of coronary heart disease are predicted and identified by collecting demographic, behavioral, and medical data from residents of the town of Framingham, Massachusetts. In general, this study selected the risk factors with high contributions to coronary heart disease as predictors by principal component analysis (PCA) and backward stepwise elimination; fitted the statistical model by performing random forest algorithm and logistic regression; and found out the more accurate results as the reference for prevention and intervention of coronary heart disease through confusion matrix and AUC-ROC curve.

Feature selection and modeling after implementing PCA and Backward stepwise elimination effectively avoids the issue of overfitting and improves the classification performance of each model. These two techniques offer assistance to medical researchers for better predicting the odds of developing CHD in statistical ways and may also be impactful on other predictable diseases. The final random forest model is optimized by grid search and used to predict the target variable by the 10 predictors from PCA results. The model comparison outcomes represent that the final random forest model has higher capacity to distinguish the ten-year future risk of coronary heart disease. However, it does not mean that other classifiers are useless. The prediction results and coefficient estimate of these classifiers can be used as references for doctors to prescribe appropriate treatment plans and preventive measures.

Nevertheless, there are several limitations in this research. Missing values and imbalance in data occurred during data collection

process may affect the reliability of the results. Moreover, the potential factors such as individual genetic differences and ten-year future death cases are not considered in this study. In the future research, more comprehensive factors and cases required to be taken into consideration.

REFERENCES

[1] Feng Y, *et al.* 2021. Adherence to antihypertensive medication and cardiovascular disease events in hypertensive patients: a dose–response meta-analysis of 2 769 700 participants in cohort study. QJM: An International Journal of Medicine 115, 5, 279-286.

[2] Jiaxin Yu, Xiaokun Liu, Shuohua Chen, Yan Liu, Hongmin Liu, Hongwei Zheng, Ning Yang, Shouling Wu, and Yuming Li. 2021. Effects of Low-density Lipoprotein Cholesterol on Cardiovascular Disease and All-cause Mortality in Elderly Patients (≥75 years old). Endocrine 75, 2, 418-426.

[3] Dariush Mozaffarian, Peter W.F. Wilson, and William B. Kannel. 2008. Beyond Established and Novel Risk Factors: Lifestyle Risk Factors for Cardiovascular Disease. Circulation (New York, N.Y.) 117, 23, 3031-3038.

[4] Lena Lönnberg, Mattias Damberg, and Åsa Revenäs. 2020. It's up to me": the experience of patients at high risk of cardiovascular disease of lifestyle change. Scandinavian Journal of Primary Health Care 38, 3, 340-351.

[5] Pamela B. Morris, *et al.* 2015. Cardiovascular Effects of Exposure to Cigarette Smoke and Electronic Cigarettes. Journal of the American College of Cardiology 66, 12, 1378-1391.

[6] Robin Nance, Joseph Delaney, John W. McEvoy, Michael J. Blaha, Gregory L. Burke, Ana Navas-Acien, Joel D. Kaufman, Elizabeth C. Oelsner, and Robyn L. McClelland. 2017. Smoking intensity (pack/day) is a better measure than pack-years or smoking status for modeling cardiovascular disease outcomes. Journal of Clinical Epidemiology 81, 111-119.

[7] Nancy A. Rigotti and Mary M. McDermott. 2019. Smoking Cessation and Cardiovascular Disease. Journal of the American College of Cardiology 74, 4, 508-511.

[8] Satoru Kodama, Shiro Tanaka, Yoriko Heianza, Kazuya Fujihara, Chika Horikawa, Hitoshi Shimano, Kazumi Saito, Nobuhiro Yamada, Yasuo Ohashi, and Hirohito Sone. 2013. Association Between Physical Activity and Risk of All-Cause Mortality and Cardiovascular Disease in Patients With Diabetes. Diabetes Care 36, 2, 471-479.

[9] Neha J. Pagidipati, Ann Marie Navar, Karen S. Pieper, Jennifer B. Green, M. Angelyn Bethel, Paul W. Armstrong, Robert G. Josse, Darren K. McGuire, Yuliya Lokhnygina, Jan H. Cornel, Sigrun Halvorsen, Timo E. Strandberg, Tuncay Delibasi, Rury R. Holman, and Eric D. Peterson. 2017. Secondary Prevention of Cardiovascular Disease in Patients With Type 2 Diabetes Mellitus. Circulation 136, 13, 1193-1203.

[10] N. A. Odunaiya, Q. A. Louw, and K. A. Grimmer. 2015. Are lifestyle cardiovascular disease risk factors associated with pre-hypertension in 15–18 years rural Nigerian youth? A cross sectional study. BMC Cardiovascular Disorders 15, 136, 144-144.

[11] Tommy Wright. 2014. Statistical Methods and the Improvement of Data Quality. Academic Press.

[12] Mohammad Zoynul Abedin, M. Kabir Hassan, Petr Hajek, and Mohammed Mohi Uddin. 2021. The Essentials of Machine Learning in Finance and Accounting. Routledge.

[13] Le Yang and Hua Chen. 2022. A Network Intrusion Detection Model Based on Principal Component Analysis and Random Forest. Frontiers in Computing and Intelligent Systems 2, 1 (2022), 76-79.

[14] AKASH JAIN. Heart Disease Prediction EDA & Classification | Kaggle. kaggle.com. Retrieved July 20, 2023 from https://www.kaggle.com/code/akash987/heart-disease-prediction-eda-classification.

[15] MICHAEL KIMOLLO. Predicting Heart Disease Risk with ML models. kaggle.com. Retrieved July 20, 2023 from https://www.kaggle.com/code/michaelkimollo/predicting-heart-disease-risk-with-ml-models.