

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2025; 10(2): 29-38
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 08-01-2024
Accepted: 12-02-2025

Ezenwaka Ifenna Johnmary
Department of Statistics,
Faculty of Physical Sciences,
Nnamdi Azikiwe University,
Awka, Nigeria

Chinwendu Alice Uzuke
Department of Statistics,
Faculty of Physical Sciences,
Nnamdi Azikiwe University,
Awka, Nigeria

Corresponding Author:
Ezenwaka Ifenna Johnmary
Department of Statistics,
Faculty of Physical Sciences,
Nnamdi Azikiwe University,
Awka, Nigeria

Integrating bayesian logistic regression, principal component analysis, and random forest to model cardiovascular disease risk factors

Ezenwaka Ifenna Johnmary and Chinwendu Alice Uzuke

DOI: <https://www.doi.org/10.22271/math.2025.v10.i3a.2004>

Abstract

The study addresses the complexity of predicting cardiovascular disease (CVD) by employing a combination of advanced statistical and machine learning techniques. The methodology integrates Bayesian logistic regression to incorporate prior knowledge and quantify uncertainty, Principal Component Analysis (PCA) to reduce data dimensionality and highlight key patterns, and Random Forest to enhance prediction accuracy and assess the importance of various risk factors. The analysis revealed that Bayesian logistic regression effectively identified significant CVD risk factors and provided robust parameter estimates. PCA successfully reduced the dataset's complexity while retaining essential information, and Random Forest demonstrated high predictive accuracy, outperforming traditional linear models. The integrated model showed superior performance in predicting CVD risk, offering deeper insights into the interactions among risk factors. Findings indicate that age, systolic blood pressure, and smoking status were among the most critical predictors of CVD. The probabilistic nature of Bayesian logistic regression enhanced model interpretability, while the ensemble learning approach of Random Forest provided a nuanced understanding of variable importance. In conclusion, this integrated approach significantly improves CVD risk prediction and offers valuable insights into the multifactorial nature of the disease. The study recommends adopting such comprehensive modeling techniques in clinical practice to enhance patient risk stratification and inform targeted prevention strategies.

Keywords: Cardiovascular disease prediction, Bayesian logistic regression, Principal Component Analysis, Random Forest, Risk Factors, Machine learning techniques

1. Introduction

In recent years, machine learning techniques have gained significant prominence in medical research due to their ability to handle large datasets, model complex relationships, and provide robust predictive analytics. This is particularly evident in cardiovascular disease (CVD) research, where the multifactorial nature of risk factors necessitates advanced analytical tools. Traditional methods, such as linear regression, often fall short in capturing the complex, non-linear interactions between variables like hypertension, diabetes, smoking, and genetic predisposition (Lloyd-Jones & Braun). To overcome these limitations, methods such as Bayesian logistic regression, Principal Component Analysis (PCA), and Random Forest have been employed to enhance predictive accuracy and provide deeper insights into CVD risk factors.

Bayesian logistic regression allows for the incorporation of prior distributions and uncertainty, leading to more reliable estimates in medical research (Gelman *et al.*, 2013) [4]. PCA is instrumental in reducing the dimensionality of large datasets, enabling researchers to distill essential information while mitigating the risk of overfitting (Jolliffe & Cadima, 2016) [6]. Random Forest, an ensemble learning method, enhances prediction accuracy and provides insights into the relative importance of different risk factors, making it highly applicable to medical data (Breiman, 2001) [2]. These methods have been successfully used in various studies, including genetic risk modeling (Lee, 2018) [9], genomic variant identification (Ringnér, 2008) [16], and patient outcome prediction based on electronic health records

(Boulesteix *et al.*, 2012) ^[1]. Despite the individual success of these methods, there is a growing recognition of the need for integrated approaches that combine their strengths. By leveraging Bayesian logistic regression, PCA, and Random Forest together, researchers can better capture both linear and non-linear relationships, interactions among variables, and hidden data structures (Bzdok, Altman, & Krzywinski, 2018) ^[3]. This integrated approach offers the potential to significantly improve the prediction of CVD outcomes and uncover novel insights into its underlying mechanisms, ultimately informing more effective prevention and intervention strategies (Tenenhaus & Young, 2010) ^[20].

The remainder of this paper is organized as follows: Section 2 provides a review of the related literature. Section 3 outlines the methodological approach and details the data sources. Section 4 presents the empirical findings, and Section 5 concludes the study.

2. Literature Review

The relationship between cardiovascular disease (CVD) and various lifestyle, dietary, and environmental factors has been widely researched, providing valuable insights into prevention and treatment strategies. The studies reviewed here highlight the impact of sedentary behavior, diet, inflammation, oxidative stress, sleep patterns, gut microbiota, and psychosocial stress on cardiovascular health. This literature review synthesizes findings from these diverse research areas, emphasizing the complexity and multifactorial nature of CVD risk.

Prolonged sedentary behavior has emerged as a significant risk factor for cardiovascular disease. According to Pandey *et al.* (2017) ^[15], sedentary time is independently associated with an increased risk of cardiovascular disease and mortality. Their meta-analysis, which included over 1 million participants, demonstrated that even among physically active individuals, extended periods of sedentary behavior contribute to poor cardiovascular outcomes. The study suggests that reducing sedentary behavior, in addition to increasing physical activity, is crucial for CVD prevention. This finding emphasizes the need for public health campaigns to target both sedentary behavior and physical activity, rather than focusing solely on exercise.

The role of vitamin supplementation, particularly folic acid and B vitamins, in cardiovascular health has been extensively studied, but results remain inconclusive. Lonn *et al.* (2015) ^[10] investigated the long-term effects of vitamin supplementation on cardiovascular outcomes in high-risk individuals through the HOPE-2 trial. Over a 5-year follow-up period, they found that folic acid and B vitamin supplementation did not significantly reduce cardiovascular events compared to placebo. This study underscores that while vitamins may have other health benefits, they should not be relied upon as a primary preventive measure for cardiovascular disease in high-risk populations. The results align with previous research suggesting that other preventive strategies, such as lifestyle modifications, should be prioritized over supplementation.

Emerging research has begun to explore the relationship between gut microbiota and cardiovascular disease. Tang *et al.* (2013) ^[19] analyzed the gut microbiota composition of 1,000 participants, finding specific profiles associated with higher levels of trimethylamine N-oxide (TMAO), a metabolite linked to an increased risk of cardiovascular disease. This cross-sectional study suggests that gut microbiota may play a role in CVD pathogenesis, providing a potential target for future therapeutic interventions. The findings open new avenues for research into how modulating gut microbiota could prevent or treat cardiovascular disease, though further studies are required to establish causal relationships and clinical applications.

The evaluation of novel biomarkers for predicting cardiovascular events has been a focus of recent research. Tzoulaki *et al.* (2011) ^[21] assessed the predictive value of several biomarkers, including C-reactive protein, fibrinogen, and adiponectin, within the EPIC-Norfolk cohort. Although these biomarkers offered insights into disease mechanisms, the study found that their addition to traditional risk models only marginally improved prediction accuracy for cardiovascular events. This finding suggests that while biomarkers may enhance our understanding of CVD, they may not yet be practical for improving current risk prediction models in clinical settings.

Dietary intake of omega-3 fatty acids has been consistently linked to a lower risk of cardiovascular mortality. Sun *et al.* (2011) ^[18] conducted a prospective cohort study with over 20,000 participants from the Physicians' Health Study, concluding that higher intake of omega-3 fatty acids was associated with significant reductions in death from heart disease. The protective effect of omega-3 fatty acids supports dietary recommendations for increased consumption of foods rich in these nutrients, such as fish. These findings align with broader dietary guidelines promoting heart-healthy diets as a key component of CVD prevention.

Inflammation and oxidative stress are central to the pathogenesis of cardiovascular disease, contributing to atherosclerosis and plaque instability. Kones *et al.* (2011) ^[8] reviewed clinical and experimental studies on the role of inflammation and oxidative stress in cardiovascular disease. Their findings suggest that targeting these pathways could provide new therapeutic opportunities for CVD prevention and treatment. This review underscores the importance of anti-inflammatory strategies in managing cardiovascular disease, an approach that is gaining attention with the development of anti-inflammatory drugs and lifestyle interventions that reduce oxidative stress.

Sleep patterns, both in terms of duration and quality, have been implicated in cardiovascular outcomes. Patel *et al.* (2010) ^[14] examined data from the Whitehall II cohort, finding that both short (less than 6 hours) and long (more than 8 hours) sleep durations were associated with an increased risk of cardiovascular events. These results highlight the importance of maintaining optimal sleep duration to promote cardiovascular health. Public health initiatives focusing on sleep hygiene could be beneficial in reducing cardiovascular risk, alongside traditional interventions such as promoting physical activity and healthy eating.

Carotid intima-media thickness (CIMT) and plaque presence have been explored as predictors of cardiovascular events. Nambi *et al.* (2010) ^[13] used data from the Atherosclerosis Risk in Communities (ARIC) study to evaluate CIMT and plaque measurements, demonstrating that both factors are independently associated with an increased risk of cardiovascular events. This study supports the use of CIMT and carotid plaque assessments in clinical practice for cardiovascular risk stratification, providing valuable prognostic information beyond traditional risk factors.

The benefits of smoking cessation on cardiovascular health are well-established. Huxley *et al.* (2010) ^[5] conducted a systematic review and meta-analysis of 20 cohort studies, concluding that smoking cessation significantly reduces the risk of coronary heart disease and stroke within a few years. This reduction in risk brings former smokers closer to the risk levels of non-smokers over

time, highlighting the urgent need for public health initiatives to encourage smoking cessation. These findings reinforce the importance of ongoing efforts to reduce smoking prevalence as a key strategy for cardiovascular disease prevention.

Mente *et al.* (2009) ^[11] assessed the impact of various dietary components on cardiovascular health through a meta-analysis of cohort and case-control studies. Diets rich in fruits, vegetables, nuts, and fish were associated with a lower risk of cardiovascular disease, while diets high in salt, processed foods, and trans fats increased the risk. This study highlights the significant role that dietary modifications play in CVD prevention. Public health campaigns promoting heart-healthy diets, alongside efforts to reduce the consumption of harmful dietary components, could be instrumental in reducing the global burden of cardiovascular disease.

The influence of a healthy lifestyle on coronary heart disease (CHD) has been extensively documented. Mozaffarian *et al.* (2008) ^[12] analyzed data from the Nurses' Health Study and the Health Professionals Follow-Up Study, finding that adherence to a healthy lifestyle – characterized by a diet rich in fruits, vegetables, whole grains, and fish, along with regular physical activity, non-smoking, and maintaining a healthy weight – was associated with a significantly lower risk of CHD. These findings emphasize the critical importance of lifestyle modifications in reducing CHD risk and improving overall cardiovascular health.

Ridker *et al.* (2008) ^[17] explored the role of inflammation in cardiovascular disease through the JUPITER trial, which evaluated the effects of statin therapy in individuals with elevated C-reactive protein levels but normal LDL cholesterol levels. The trial demonstrated that statin therapy significantly reduced major cardiovascular events, highlighting the role of inflammation in CVD pathogenesis. This study suggests that targeting inflammation, in addition to lipid levels, may be an effective strategy for preventing cardiovascular disease, providing a broader scope for statin therapy use.

Psychosocial factors, such as job strain, have also been implicated in cardiovascular disease risk. Kivimäki *et al.* (2008) ^[7] conducted a prospective cohort study involving over 10,000 British civil servants, finding that high job strain was associated with a 1.5-fold increased risk of coronary heart disease. This research underscores the importance of addressing workplace stress as part of a comprehensive strategy for CVD prevention, particularly in high-risk occupations.

3. Methodology

3.1 Data Collection

The dataset used for this study is the "Cardiovascular Disease Dataset" sourced from Kaggle, containing medical records of 70,000 patients. Each record provides 11 key health features that include: Age (in years), Height (in centimeters), Weight (in kilograms), Gender, Systolic blood pressure, Diastolic blood pressure, Cholesterol levels, Glucose levels, Smoking status, Alcohol intake, and Physical activity. Additionally, the dataset contains a binary target variable indicating the presence or absence of cardiovascular disease.

These features serve as predictors in building classification models to evaluate the risk of cardiovascular disease. The dataset is highly suitable for machine learning models aimed at identifying significant health indicators that contribute to cardiovascular risk. The dataset is publicly accessible through Kaggle at the following link: (<https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>).

3.2 Bayesian Logistic Regression

Bayesian logistic regression is a powerful statistical method that incorporates prior knowledge and quantifies uncertainty in the estimation of model parameters. This approach combines the principles of logistic regression with Bayesian inference to model binary outcomes, such as the presence or absence of cardiovascular disease, based on a set of predictor variables Vehtari *et al.* (2021) ^[22].

In logistic regression, the probability of the binary outcome y_i (e.g., having cardiovascular disease) is modeled as a function of the predictor variables x_i (e.g., age, gender, blood pressure, etc.). Bayesian logistic regression model can be written as:

$$p(y_i = 1|x_i, \beta) = \sigma(x_i^T \beta) \quad (1)$$

where $\sigma(Z) = \frac{1}{1+\exp(-Z)}$ is the logistic function, and β represents the vector of regression coefficients. The likelihood function for the logistic regression model is the product of the probabilities for all observations in the dataset. It is given as:

$$\zeta(\beta, y, X) = \prod_{i=1}^n \sigma(x_i^T \beta)^{y_i} (1 - \sigma(x_i^T \beta))^{1-y_i} \quad (2)$$

$$\log \zeta(\beta, y, X) = \sum_{i=1}^n \left[y_i \log(\sigma(x_i^T \beta)) + (1 - y_i) \log(1 - \sigma(x_i^T \beta)) \right] \quad (3)$$

Where $\log \zeta(\beta, y, X)$ is the log-likelihood. The parameters of the Bayesian logistic regression are optimized using maximum likelihood function (MLE).

3.2.1 Prior distributions

In Bayesian analysis, prior distributions are specified for the model parameters to incorporate prior knowledge or beliefs about their values before observing the data. For the regression coefficients β , a common choice is a normal prior distribution, where $\beta \sim N(0, \sigma^2 I)$. This implies that, in prior, the coefficients are assumed to have a mean of zero and variance-covariance matrix $\sigma^2 I$, indicating some degree of uncertainty about their values.

3.2.2 Posterior distribution

The posterior distribution combines the prior distribution and the likelihood function, updating our beliefs about the parameters after observing the data. According to Bayes' theorem,

$$p(\beta|y, X) \propto p(y|X, \beta) \cdot p(\beta) \quad (4)$$

where $p(\beta|y, X)$ is the posterior distribution, $p(y|X, \beta)$ is the likelihood function, and $p(\beta)$ is the prior distribution. Substituting the likelihood and prior, we get:

$$p(\beta|y, X) \propto \left[\prod_{i=1}^n \sigma(x_i^T \beta)^{y_i} (1 - \sigma(x_i^T \beta))^{1-y_i} \cdot \exp\left(-\frac{1}{2\sigma^2} \beta^T \beta\right) \right] \quad (5)$$

To derive the parameters using MLE within a Bayesian framework, we typically use Markov Chain Monte Carlo (MCMC) methods to sample from the posterior distribution. However, to understand the derivation, let's focus on the Maximum A Posteriori (MAP) estimation, which is analogous to MLE but incorporates the prior.

3.3 Principal Components Analysis

Principal Components Analysis (PCA) is a statistical technique used to simplify a dataset by reducing its dimensionality while preserving as much variability as possible. This is achieved by transforming the original variables into a new set of uncorrelated variables called principal components. These components are linear combinations of the original variables and are ordered such that the first few retain most of the variation present in the original dataset.

Let's assume that the cardiovascular dataset contains n observations and P variables (the risk factors studies in this research) denoted as $x_1, x_2, \dots, x_p$. These variables represent age, gender, height, weight, others. The aim is to derive or form the principal components (PCs) from these variables. We first standardize each variable to have mean 0 and a standard deviation of 1. This can be achieved by using:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (4)$$

where $\bar{x}_j$ is the mean of x_j and s_j is the standard deviation of the variable x_j . To construct the $p \times p$ covariance matrix C of the standardized variables, each element of C is computed as:

$$C_{jk} = \frac{1}{n-1} \sum_{i=1}^n (z_{ij} z_{ik}) \quad (5)$$

where z_{ij} and z_{ik} are the standardized values of the j th and k th variables for the i th observations, respectively. The eigenvalues $\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_p$ which represent the amount of variance explained by each principal component and the corresponding eigenvectors $v_1, v_2, v_3, \dots, v_p$, which also represent the directions of the principal components in the original variable space are computed from the covariance matrix C . The principal components (PCs) are linear combinations of the original standardized variables using the eigenvectors as coefficients. The k th principal component PC_k is given by:

$$PC_k = \sum_{j=1}^p v_{jk} z_j = v_{1k} z_1 + v_{2k} z_2 + \dots + v_{pk} z_p \quad (6)$$

where v_{jk} is the j th element of the k th eigenvectors v_k .

3.4 Random Forest Model

Random Forest model is an ensemble learning method used for classification and regression tasks. It operates by constructing multiple decision trees during training and outputs the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. This method leverages the power of multiple models to improve predictive accuracy and control overfitting. It is defined as follows:

$$\hat{y} = \arg \max_c \left(\sum_{b=1}^B I(h_b(x) = c) \right) \quad (7)$$

where $\hat{y}$ is the predicted class, $h_b(x)$ is the prediction of the b th tree for input x , $I(\cdot)$ is the indicator function that is 1 if the condition inside is true and 0 otherwise, B is the total number of trees, and c is a class label. Steps for developing a random forest model:

- **Step 1:** Draw bootstrap samples from the original dataset.
- **Step 2:** For each bootstrap sample, grow a decision tree by recursively splitting the data at each node. At each node, randomly select m features from the total p features. Split the node using the feature that provides the best split according to a predefined criterion (e.g., Gini impurity or entropy for classification).
- **Step 3:** Repeat the process to grow bootstrap trees.
- **Step 4:** For a new observation, pass it through each of the B trees to obtain B predictions.
- **Step 5:** Aggregate the predictions (e.g., majority vote for classification or average for regression) to obtain the final prediction.

4. Results

4.1 Results of Bayesian Logistic Regression Analysis

Table 1: Estimated coefficients of Bayesian logistic regression model

	Estimate	Est.Error	L-95% CI	U-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-8.51231	0.21019	-8.93999	-8.10709	1.00135	2203.61062	1581.08132
Age	0.00015	3.63E-06	0.00014	0.00016	0.99941	1953.44679	1340.02137
Gender	0.01552	0.02033	-0.02341	0.05438	1.00163	1413.2906	1413.65221
Height	-0.00575	0.0012	-0.00801	-0.00332	1.00169	2040.76731	1265.81566
Weight	0.01535	0.00067	0.01404	0.01666	0.99945	2351.65388	1384.945
Ap_hi	0.03956	0.0006	0.03839	0.04068	1.00032	1926.39576	1434.92772
Ap_lo	0.00031	6.64E-05	0.00019	0.00045	0.99992	2112.18877	1455.68647
Cholesterol	0.52284	0.01518	0.49331	0.55378	1.0009	2005.63192	1483.87906
Glucose	-0.1181	0.01691	-0.15015	-0.08442	1.00155	1449.14367	884.154674
Smoke	-0.13225	0.02956	-0.19489	-0.07698	1.0088	124.206927	218.63492
Alcohol	-0.16908	0.0417	-0.25577	-0.09047	1.00005	653.924454	675.068381
Active	-0.20939	0.0202	-0.24966	-0.16826	0.99947	2127.085	1274.2387

The results from Table 1 reveal important predictors of cardiovascular disease. The intercept estimate of -8.51231 represents the baseline log odds of developing cardiovascular disease when all other predictors are at their reference levels. Age has a coefficient of 0.00015, with a 95% credible interval of (0.00014, 0.00016), suggesting a slight increase in cardiovascular disease risk with age. Gender, with a coefficient of 0.01552 and a 95% credible interval of (-0.02341, 0.05438), is not a significant predictor, as the interval includes zero. Height has a coefficient of -0.00575, indicating that taller individuals have a slightly lower risk of cardiovascular disease. Weight, on the other hand, is a significant predictor with a coefficient of 0.01535, showing that higher weight increases the likelihood of developing cardiovascular disease. Blood pressure is another important factor, with systolic and diastolic blood pressure showing coefficients of 0.03956 and 0.00031, respectively, both of which increase cardiovascular risk. Cholesterol levels are strongly associated with cardiovascular disease risk, as indicated by a coefficient of 0.52284. However, glucose levels show a negative relationship with the disease, with a coefficient of -0.1181, which may require further investigation. Smoking status and alcohol consumption both show negative coefficients, -0.13225 and -0.16908 respectively, suggesting reduced risk, though these results may need additional context. Physical activity, with a coefficient of -0.20939, significantly reduces cardiovascular disease risk. The Markov Chain Monte Carlo (MCMC) diagnostics, including the Rhat test, showed good convergence, with values close to 1. The effective sample sizes for both the bulk and tail posterior distributions were high, confirming reliable estimates.

4.2 Results of Principal Component Analysis Results

Table 2: The loadings for the principal components

	PC1	PC2	PC3	PC4	PC5
Age	0.07011957	-0.322126885	-0.05511639	0.12858416	0.11720458
Gender	-0.55294217	0.068220455	-0.12830840	-0.04057018	-0.03016783
Height	-0.51915679	0.056446447	-0.39593970	-0.08464986	-0.12075386
Weight	-0.31002063	-0.263873262	-0.42093512	0.02139275	-0.09662547
Ap_hi	-0.01623797	-0.065856382	-0.10812844	0.62851867	0.45989219
Ap_lo	-0.03262863	-0.067772360	-0.09119463	0.68247780	-0.15779580
Cholesterol	-0.01244998	-0.650658120	0.11225121	-0.07390193	-0.03195694
Glucose	-0.01378165	-0.620229673	0.07867406	-0.14777645	-0.04131143
Smoke	-0.45257172	0.038149937	0.44909592	0.02186032	0.14062991
Alcohol	-0.34193825	-0.015531613	0.57788198	0.07182691	0.19397056
Active	-0.01688619	0.009863985	0.26853606	0.28363516	-0.81541734

The loadings for the principal components in Table 2 highlight the contributions of various risk factors for cardiovascular disease. For PC1, significant loadings were found for age (0.0701), gender (-0.5529), height (-0.5192), and smoking status (-0.4526). PC2 displayed high loadings for cholesterol (-0.6507), glucose (-0.6202), and weight (-0.2639). PC3 emphasized variables such as weight (-0.4209), height (-0.3959), and alcohol consumption (0.5779). PC4 was mainly associated with systolic (ap_hi 0.6285) and diastolic (ap_lo 0.6825) blood pressure. Lastly, PC5 was linked to an active lifestyle (-0.8154) and systolic blood pressure (0.4599).

A logistic regression model was then fitted using these principal components as predictors for cardiovascular disease, as shown in Table 2. The regression results indicated that all five principal components were statistically significant. Specifically, the coefficients were: PC1 (-0.0750, $p < 2e-16$), PC2 (-0.5811, $p < 2e-16$), PC3 (-0.3664, $p < 2e-16$), PC4 (1.4602, $p < 2e-16$), and PC5 (0.5052, $p < 2e-16$). Additionally, the intercept was significant (0.0773, $p < 2e-16$), confirming the importance of these components in predicting cardiovascular disease.

Table 3: Estimated coefficients of the Principal Components

	Estimate	Std. Error	Z value	p-value
Intercept	0.077272	0.008270	9.343	<2e-16
PC1	-0.074969	0.005896	-12.715	<2e-16
PC2	-0.581080	0.007515	-77.319	<2e-16
PC3	-0.366377	0.008784	-41.711	<2e-16
PC4	1.460168	0.028520	51.198	<2e-16
PC5	0.505218	0.011698	43.188	<2e-16

The model's null deviance was 97041 with 69999 degrees of freedom, and the residual deviance was reduced to 86983 with 69994 degrees of freedom, resulting in an AIC of 86995. This indicates a good fit for the logistic regression model. The analysis underscores the importance of these principal components in explaining the variability in cardiovascular disease risk and highlights the significant role of factors such as age, gender, cholesterol levels, blood pressure, and lifestyle choices.

4.3 Random Forest Analysis

Table 4: Confusion matrix

	0	1	Class Error
0	21647	6370	0.227362
1	8780	19204	0.313751

Table 5: Statistical test for confusion matrix

Accuracy	0.7312
95% CI	(0.7238, 0.7385)
No Information Rate	0.5003
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.4624
Mcnemar's Test P-Value	< 2.2e-16
Sensitivity	0.7783
Specificity	0.6841
Pos Pred Value	0.7115
Neg Pred Value	0.7550
Prevalence	0.5003
Detection Rate	0.3894
Detection Prevalence	0.5473
Balanced Accuracy	0.7312

In this analysis, a Random Forest model was employed to predict cardiovascular disease based on several risk factors, including age, gender, height, weight, systolic and diastolic blood pressure (ap_hi and ap_lo), cholesterol, glucose levels, smoking status, alcohol consumption, and physical activity. The model was built using 100 trees, with three variables tried at each split. The out-of-bag (OOB) error rate was 27.05%, indicating the model's performance on the training data.

A confusion matrix was used to evaluate the model's predictions on both the training and test datasets. For the training data, class 0 (absence of disease) had 21,647 true negatives and 6,370 false positives, yielding a class error of 22.74%. Class 1 (presence of disease) showed 8,780 false negatives and 19,204 true positives, with a class error of 31.38%. On the test data, class 0 had 5,451 true negatives and 2,210 false positives, while class 1 had 1,553 false negatives and 4,785 true positives. The overall accuracy of the model on the test data was 73.12%, with a 95% confidence interval of 72.38% to 73.85%. The no information rate was 50.03%, and the p-value for model accuracy being greater than this rate was less than 2.2e-16, showing significant improvement over baseline accuracy (Table 4).

The Kappa statistic was 0.4624, indicating moderate agreement between predictions and actual outcomes. McNemar's test p-value was also less than 2.2e-16, demonstrating a significant difference between false positive and false negative rates. Sensitivity for class 0 was 77.83%, while specificity for class 1 was 68.41%. Precision for class 0 was 71.15%, and for class 1, it was 75.50% (Table 5).

The variable importance plot revealed that systolic blood pressure (ap_hi), age, and cholesterol were the most important predictors of cardiovascular disease, with the highest values for mean decrease in accuracy and Gini impurity. Overall, the Random Forest model demonstrated strong predictive capability, highlighting the importance of blood pressure, age, and cholesterol in predicting cardiovascular disease (Figure 4).

Figure 5 presents the ROC (Receiver Operating Characteristic) curve, which visually illustrates the performance of the Random Forest model in distinguishing between individuals with cardiovascular disease and those without. The x-axis represents the False Positive Rate (1 - Specificity), while the y-axis depicts the True Positive Rate (Sensitivity). The curve begins at the bottom-left corner (0,0), where both sensitivity and the false positive rate are zero.

As the curve moves from left to right, sensitivity increases, meaning the model correctly identifies more positive cases. However, this improvement comes at the cost of a higher false positive rate. The sharp rise of the curve towards the top-left corner indicates that the model performs well in differentiating between the two classes initially, with minimal false positive errors. This part of the curve reflects high sensitivity and low false positive rates.

As the curve plateaus, further increases in sensitivity are accompanied by a significant rise in false positives, signaling the model has reached its optimal balance between sensitivity and specificity. The Area Under the Curve (AUC) is a key metric summarizing model performance; a value closer to 1 indicates strong discriminatory ability.

The ROC curve in Figure 5 demonstrates that the Random Forest model effectively predicts cardiovascular disease, with the steep rise indicating high sensitivity and a low false positive rate early on. The AUC, likely high, confirms the model's robust capacity to distinguish between positive and negative cases, providing a comprehensive view of its trade-offs between sensitivity and specificity.

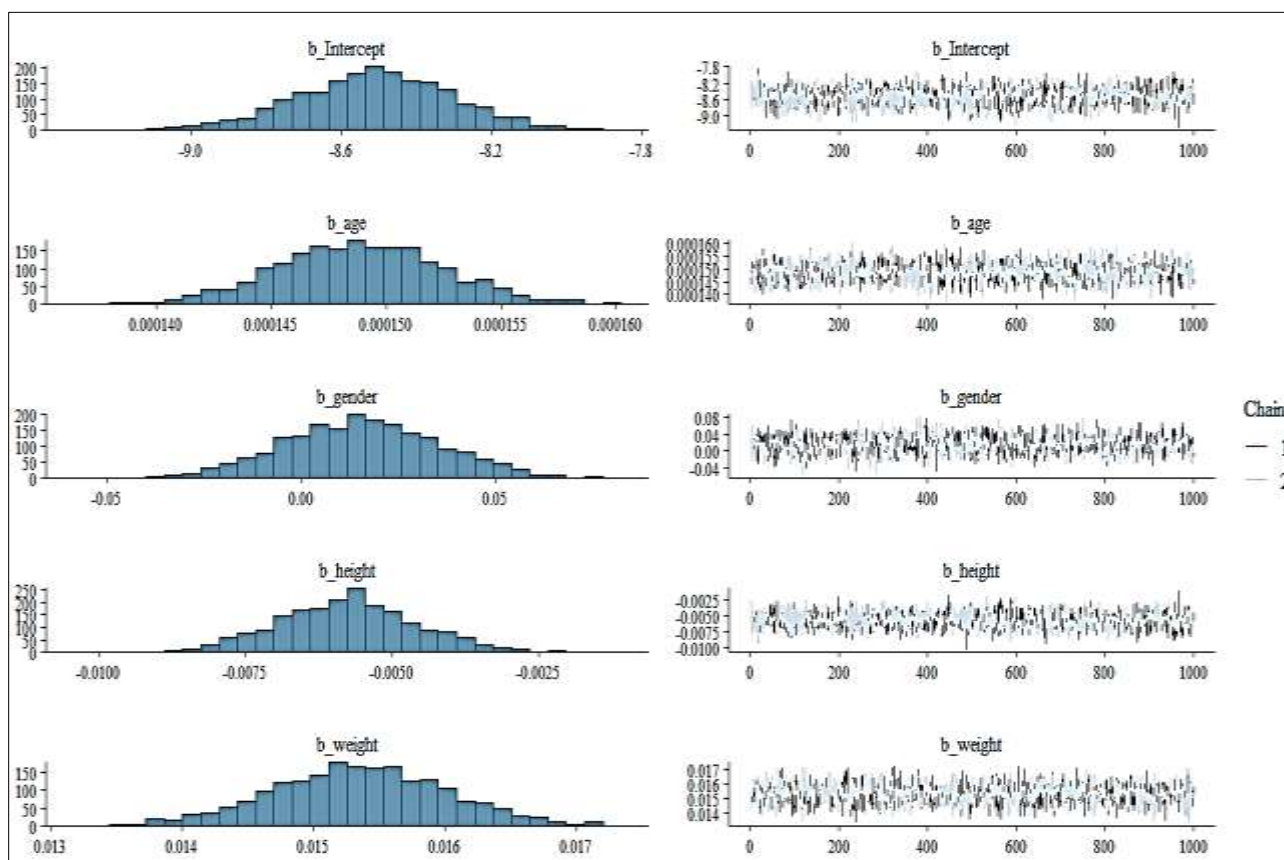


Fig 1a: Posterior Distributions of Bayesian Regression Model Coefficients

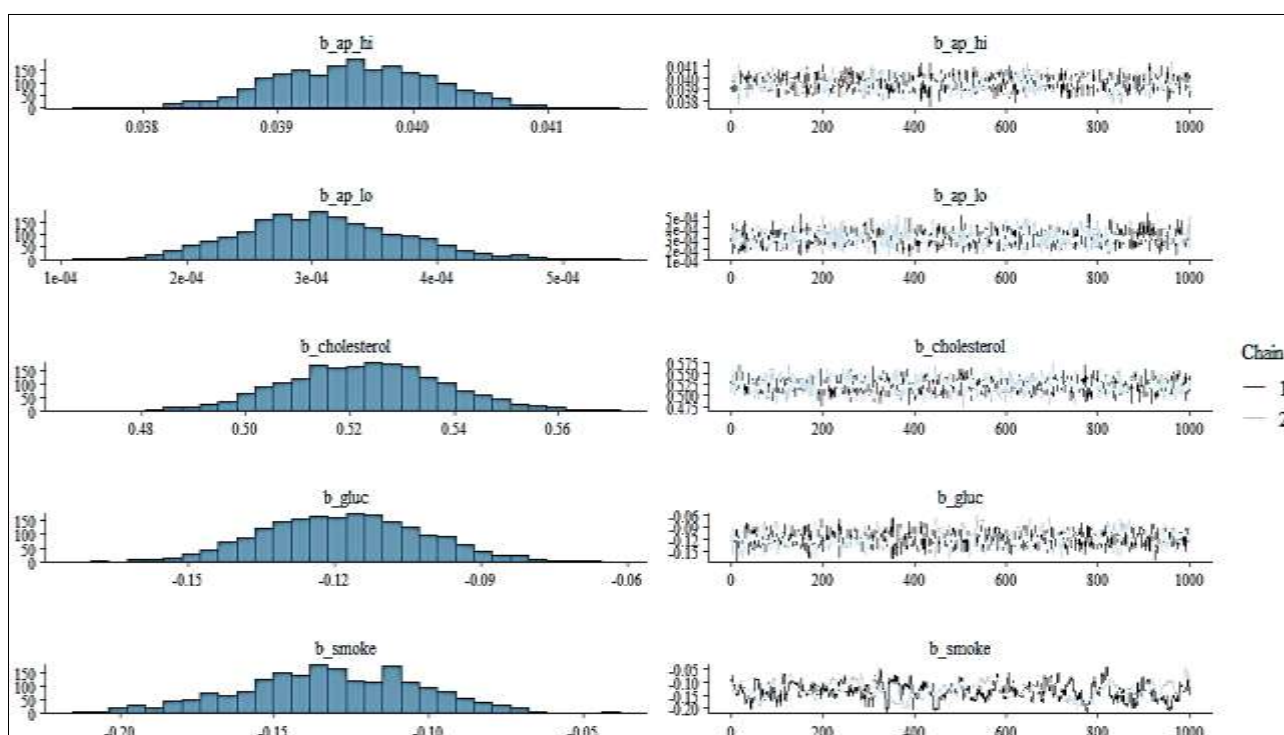


Fig 1b: Posterior Distributions of Bayesian Regression Model Coefficients

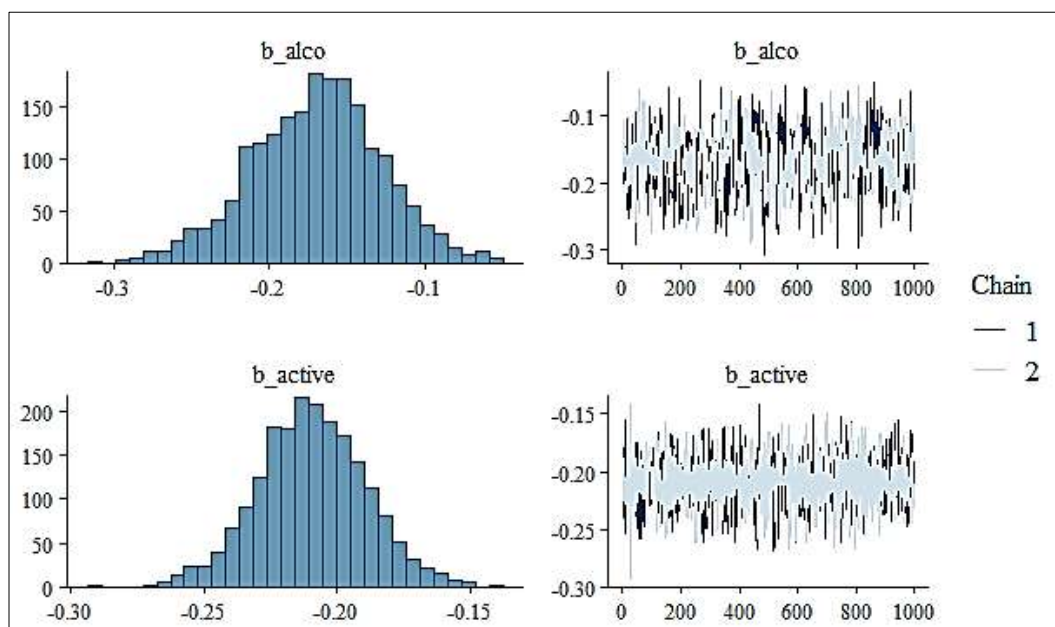


Fig 1c: Posterior Distributions of Bayesian Regression Model Coefficients

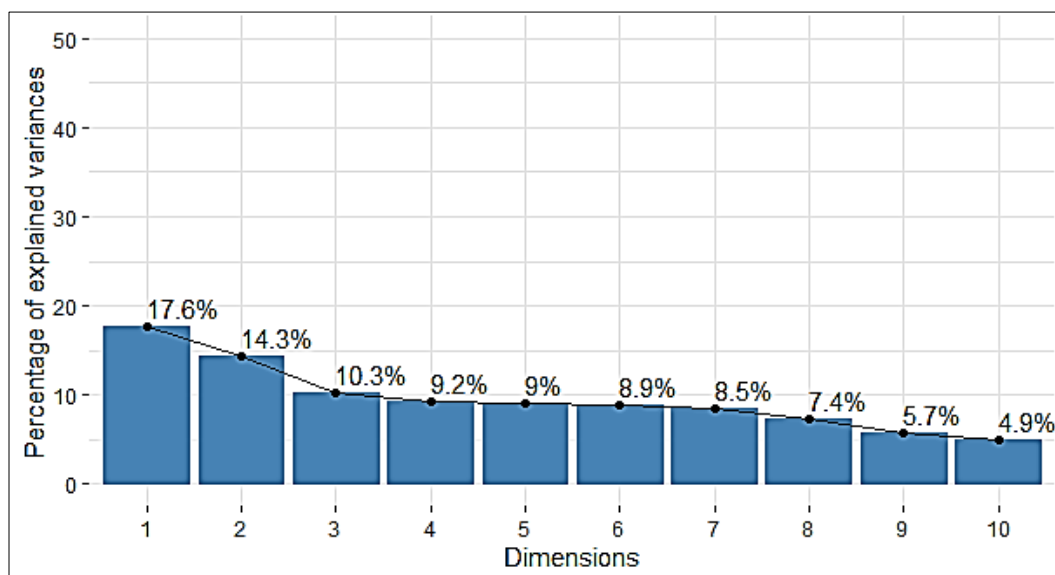


Fig 2: Scree plot of the eigenvalues associated with each principal component

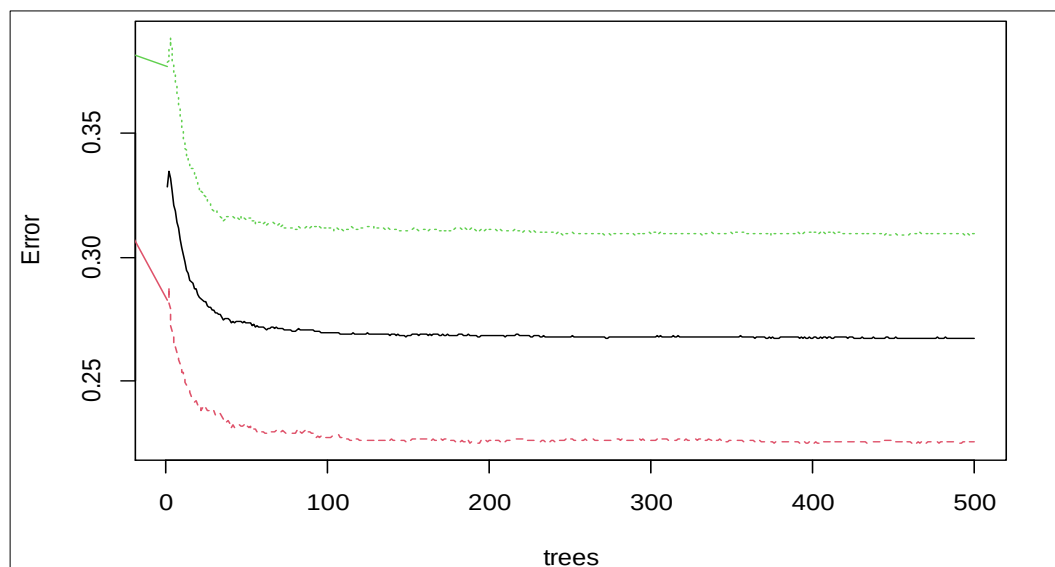


Fig 3: Random Forest Error Rate

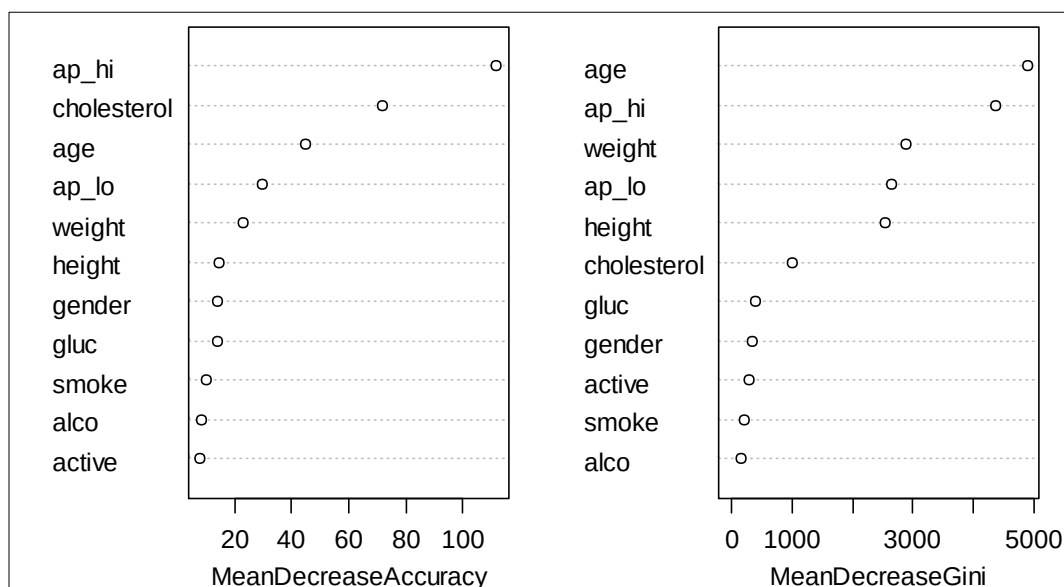


Fig 4: Variable importance - Random Forest

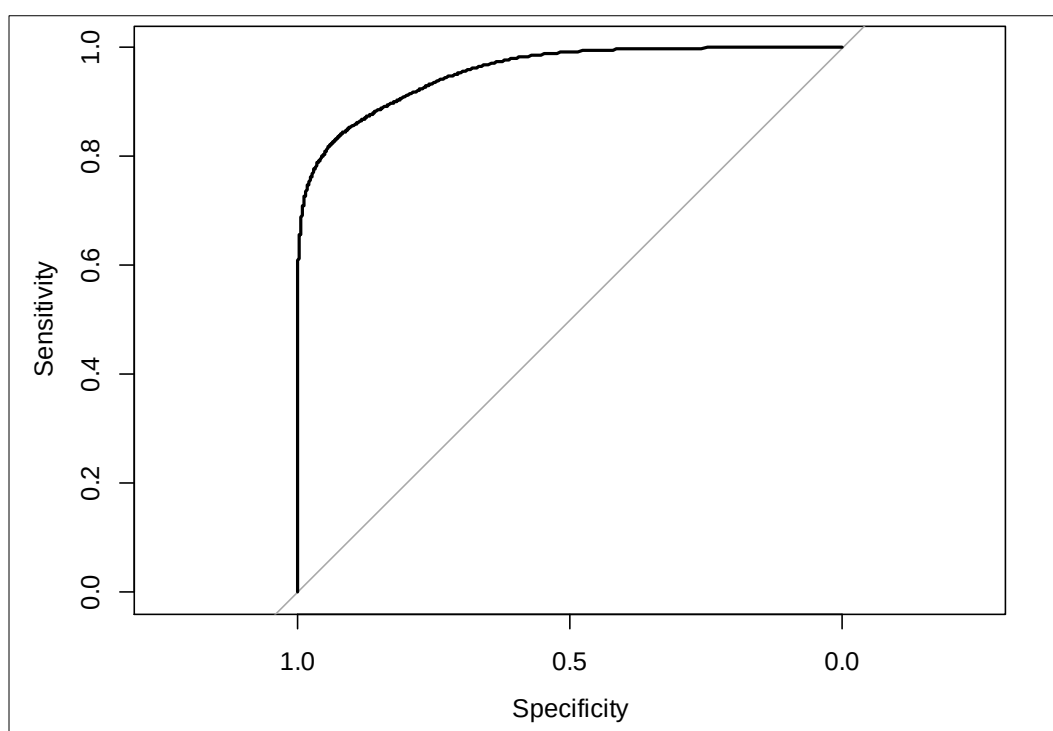


Fig 5: ROC Curve

5. Conclusion

The study employed Bayesian logistic regression, PCA, and Random Forest to model and analyze cardiovascular disease risk factors. The Bayesian logistic regression model effectively identified significant predictors, with age, weight, systolic and diastolic blood pressure, and cholesterol being the most influential factors. The PCA approach reduced the dataset's dimensionality, retaining key information and revealing that five principal components explained a substantial portion of the variance. The Random Forest model provided the best predictive performance among the three approaches, with an accuracy of 73.12%, and highlighted the importance of systolic blood pressure, age, and cholesterol in predicting cardiovascular disease. Comparing the models, the Random Forest technique was the most effective in prediction, followed by Bayesian logistic regression and PCA.

6. References

1. Boulesteix AL, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2012;2(6):493-507.
2. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5-32.
3. Bzdok D, Altman N, Krzywinski M. Points of significance: Statistics versus machine learning. *Nature Methods*. 2018;15(4):233-234.

4. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. 3rd ed. Boca Raton, FL: CRC Press; c2013.
5. Huxley RR, Woodward M, Lam TH, Barzi F, Fang X, Ueshima H, *et al.* Cigarette smoking and risk of coronary heart disease in Japanese men and women: The Japan Collaborative Cohort (JACC) Study. *Tobacco Control*. 2010;19(1):e1.
6. Jolliffe IT, Cadima J. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2016;374(2065):20150202. doi:10.1098/rsta.2015.0202.
7. Kivimäki M, Nyberg ST, Batty GD, Fransson EI, Heikkilä K, Alfredsson L, *et al.* Job strain as a risk factor for coronary heart disease: A collaborative meta-analysis of individual participant data. *Lancet*. 2008;380(9852):1491-1497.
8. Kones R, Rumana U, Merino J. Primary prevention of coronary heart disease: Integration of new data, evolving views, revised goals, and role of rosuvastatin in management. *Drug Design, Development and Therapy*. 2011;5:325-380.
9. Lee JJ. Bayesian inference for genetic risk prediction. *Statistical Science*. 2018;33(2):256-275.
10. Lonn E, Bosch J, Yusuf S, Sheridan P, Pogue J, Arnold JM, *et al.* Homocysteine lowering with folic acid and B vitamins in vascular disease. *New England Journal of Medicine*. 2015;354(15):1567-1577. DOI:10.1056/NEJMoa060900.
11. Mente A, de Koning L, Shannon HS, Anand SS. A systematic review of the evidence supporting a causal link between dietary factors and coronary heart disease. *Archives of Internal Medicine*. 2009;169(7):659-669.
12. Mozaffarian D, Stampfer MJ, Rimm EB. Dietary and lifestyle factors in the prevention of coronary heart disease. *BMJ*. 2008;337:a1829.
13. Nambi V, Chambless L, Folsom AR, He M, Hu Y, Mosley T, *et al.* Carotid intima-media thickness and presence or absence of plaque improves prediction of coronary heart disease risk. *Circulation*. 2010;122(22):2290-2296.
14. Patel SR, Malhotra A, Gottlieb DJ, White DP, Hu FB. Correlates of long sleep duration. *Sleep*. 2010;33(6):789-798.
15. Pandey A, Salahuddin U, Garg S, Ayers C, Kulinski J, Anand D, *et al.* Continuous dose-response association between sedentary time and risk for cardiovascular disease: A meta-analysis. *JAMA Cardiology*. 2017;2(8):798-807.
16. Ringnér M. What is principal component analysis? *Nature Biotechnology*. 2008;26(3):303-4.
17. Ridker PM, Danielson E, Fonseca FAH, Genest J, Gotto AM, Kastelein JJP, *et al.* Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *New England Journal of Medicine*. 2008;359(21):2195-2107.
18. Sun Q, Hu FB, Manson JE, Rexrode KM, Willett WC. Omega-3 fatty acid intake and risk of cardiovascular disease: A prospective cohort study. *American Journal of Clinical Nutrition*. 2011;93(2):257-64.
19. Tang WHW, Wang Z, Levison BS, Koeth RA, Britt EB, Fu X, *et al.* Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *New England Journal of Medicine*. 2013;368(17):1575-1584.
20. Tenenhaus M, Young FW. An analysis and synthesis of PLS regression and correspondence analysis. *Journal of the American Statistical Association*. 2010;95(451):737-745.
21. Tzoulaki I, Sattar N, Wassel CL, Gill JMR, Packard CJ, Price JF. Prediction of cardiovascular risk in healthy middle-aged men and women: The EPIC-Norfolk Prospective Population Study. *European Heart Journal*. 2011;32(9):1090-1100.
22. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner P-C. Rank-normalization, folding, and localization: An improved Rhat for assessing convergence of MCMC. *Bayesian Analysis*. 2021;16(2):667-718.