Predicting heart disease using GLM, Decision Tree and SVM models

Reza Heidari

Abstract:

In the United States, heart disease is the leading cause of death, affecting various groups. Detecting and preventing heart disease risk factors is essential for minimizing the disease's impact on public health. This project focuses on using machine learning techniques, including Generalized Linear Model (GLM), Support Vector Machine (SVM), and Decision Tree, on predicting heart disease based on a dataset obtained from the Behavioral Risk Factor Surveillance System (BRFSS) of the Centers for Disease Control and Prevention (CDC). The data set consists of 319,795 rows and 18 columns, including demographic, health status, and risk factor variables. The performance of the models is measured using Confusion Matrix, accuracy, and ROC curve evaluation metrics. Contributing to an understanding of risk factors and their effect on public health, the findings of this study can aid in the development of effective heart disease prevention strategies by medical professionals.

1. Introduction

Heart disease is a significant health concern in the United States, ranking among the leading causes of death for various racial groups, including African Americans, American Indians and Alaska Natives, and white people, according to the Centers for Disease Control and Prevention (CDC). Risk factors for heart disease include high blood pressure, high cholesterol, smoking, diabetes, obesity (high BMI), a sedentary lifestyle, and excessive alcohol consumption. Detecting and preventing these risk factors is essential for reducing the impact of heart disease on individuals and communities [4].

Advancements in computational methods have enabled the application of machine learning techniques to analyze data and identify patterns that can predict a patient's risk of heart disease. In this project, we will utilize three machine learning techniques, namely Generalized Linear Model (GLM), Support Vector Machine (SVM), and Decision Tree, to predict heart disease using a dataset obtained from the Behavioral Risk Factor Surveillance System (BRFSS) of the CDC. The BRFSS conducts annual telephone surveys to collect data on the health status of U.S. residents, making it the world's largest continuously conducted health survey system [4].

The primary objective of this report is to utilize GLM, SVM, and Decision Tree models on the dataset and evaluate their performance in predicting heart disease. By identifying the significant variables that influence the likelihood of heart disease, we aim to contribute to the understanding of heart disease risk factors and provide insights for healthcare professionals to develop effective preventive strategies. Detecting and preventing heart disease risk factors is crucial for minimizing the disease's impact on public health, and therefore the findings of this study may have significant implications for healthcare practice.

2. Data

This report's dataset is derived from the CDC's Behavioral Risk Factor Surveillance System (BRFSS), which compiles information on the health status of U.S. residents through annual telephone surveys. The dataset consists of 319,795 rows and 18 columns, with queries about factors that may influence the likelihood of heart disease posed to respondents. "HeartDisease" is regarded as a binary variable with two classes: "Yes" indicates that the respondent has heart disease, and "No" indicates that the respondent does not have heart disease [4].

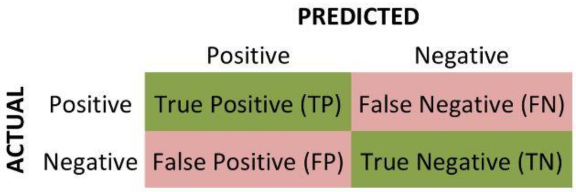
Nine Boolean variables, five string variables, and four decimal variables comprise the 18 variables in the dataset. The dataset contains the following variables: "HeartDisease", "BMI" (Body Mass Index), "Smoking" (Smoking status), "AlcoholDrinking" (Alcohol consumption), "Stroke" (History of stroke), "PhysicalHealth" (Self-reported physical health status), "MentalHealth" (Self-reported mental health status), "DiffWalking" (Difficulty walking or climbing stairs), "Sex" (Sex of the respondent), ‘AgeCategory’, "Race" (Race of the respondent), "Diabetic" (Diabetic status), ‘PhysicalActivity’ (Physical activity level), ‘GenHealth’ (General health status), ‘SleepTime’ (Sleep duration), ‘Asthma’ (History of asthma), ‘KidneyDisease’ (History of kidney disease), and ‘SkinCancer’ (History of skin cancer).

3. Method

All the R codes for implementing and analysis of the data set can be found in the following link: https://github.com/heidarir88/DASC-5420-Final-Project.git

3.1. Evaluation Metrics:

3.1.1. Confusion Matrix:



		PREDICTED	
		Positive	Negative
ACTUAL	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Figure 1: Confusion Matrix

The confusion matrix (Figure 1) is a standard visualization tool utilized in machine learning to assess the performance of a classification model. It is a square matrix that displays the number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions in a classification model made on a set of data points. Here is a list of the definitions for each part of a confusion matrix [1]:

True Positive (TP): The number of instances that are truly positive and are correctly predicted as positive by the model.

False Positive (FP): The number of instances that are actually negative but are incorrectly predicted as positive by the model.

True Negative (TN): The number of instances that are truly negative and are correctly predicted as negative by the model.

False Negative (FN): The number of instances that are actually positive but are incorrectly predicted as negative by the model.

3.1.2. Accuracy:

It represents the percentage of correctly classified instances out of the total number of instances in the dataset. The formula to calculate accuracy is [1]:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

3.1.3. AUC-ROC Curve:

The Area Under the Receiver Operating Characteristic (AUC-ROC) curve is a graphical representation of a classification model's performance. The AUC-ROC curve is generated by plotting the True Positive Rate (TPR), calculated as TP / (TP + TN), against the False Positive Rate (FPR), calculated as FP / (FP + TN), at various classification thresholds. With values ranging from 0 to 1 on both axes, the curve illustrates the trade-off between the true positive rate and the false positive rate [1].

To predict heart disease in the given dataset, I utilized three different methods: Support Vector Machine (SVM), Generalized Linear Model (GLM), and Decision Tree. To ensure fair evaluation of the models, I divided the dataset into a train set (80%) and a test set (20%).

3.1.4. GLM Method

The data was preprocessed by transforming categorical variables into factor variables. The 'AgeCategory' string variable was replaced the 'AgeCategory' string variable with the factor variables representing the average of each period. The 'Race', 'Diabetic', and 'GenHealth' variables were transformed into dummy variables. Additionally, 'BMI', 'PhysicalHealth', 'MentalHealth', and 'AgeCategory' were normalized factor variables. After implementing GLM, some predictors which were not significant in the model were removed. Mathematically, for the GLM model, the relationship can be written as

$$p(x) = \frac{e^{\beta_0 + \beta_1 X + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X + \cdots + \beta_p X_p}}$$

Where response variable p is the probability of heart disease for a person and the value between zero and one. And Xp represents the pth predictor and βp is the pth coefficient. Since the heart disease should be one or zero, the following function is applied.

$$f(x) = \begin{cases} 1\ if\ p(x) > 0.5 \\ 0\ if\ p(x) < 0.5 \end{cases}$$

3.1.5. SVM Method

All boolean and string variables were transformed into factor variables. Then SVM method was implemented with a radial kernel and the value of cost is equal to ten. In general, in classification, SVM finds the optimal hyperplane (the following formula) that best separates the data into different classes.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Where p is dimensions a hyperplane which p is equal to 2 that mean hyperplane is a line and if $\beta_0 = 0$ means the hyperplane pass the origin and the vector $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ is named the normal vector [2].

3.1.6. Decision Tree

In this project, using utilized cross-validation to determine the optimal number of terminal nodes in the decision tree, revealing that 5 terminal nodes yielded the best performance [3]. Finally, using the 'plot' function to create a plot of the size of the terminal nodes versus the deviance. (Figure 2)

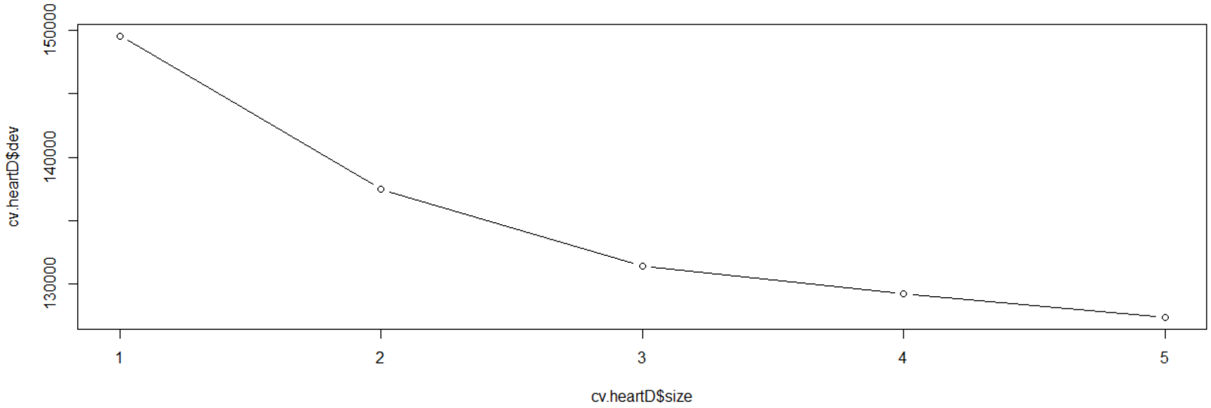


Figure 2: Plotting cv.tree Decision Tree Method

All Boolean and string variables were transformed into factor variables. The decision tree method was implemented with five terminal nodes, which was determined as the best number of terminal nodes through cross-validation. The decision tree output is shown in figure 3.

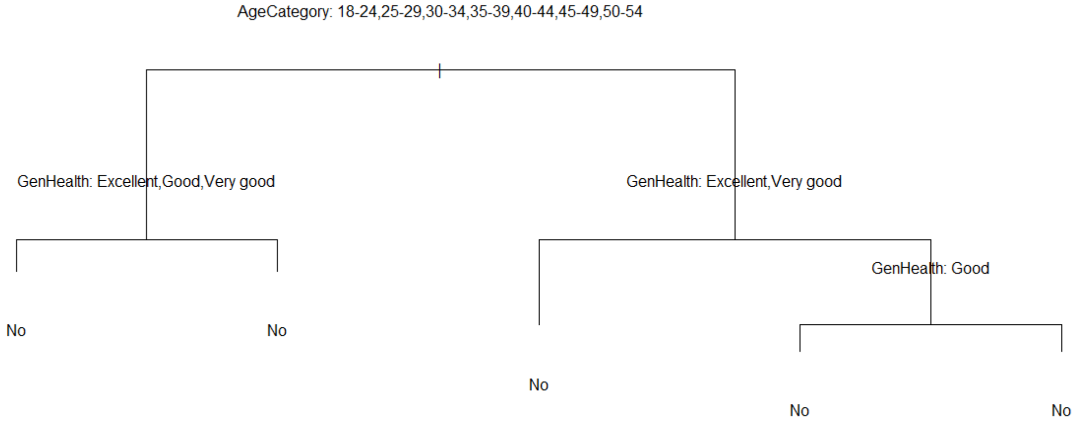

AgeCategory: 18-24,25-29,30-34,35-39,40-44,45-49,50-54
GenHealth: Excellent,Good,Very good
GenHealth: Excellent,Very good
GenHealth: Good
No
No
No
No
No

Figure 3: Plotting the Decision Tree Method

After implementing the three methods on the train data set, the three methods were evaluated their performance using three evaluation metrics: Confusion Matrix, Accuracy, and AUC-ROC Curve.

4. Results

It should be noted that the classes of the target variable, "HeartDisease", are heavily unbalanced, with potential implications for model training and evaluation (Figure 4). By generating synthetic examples of the minority class through oversampling to address class imbalance in binary classification in this data set.

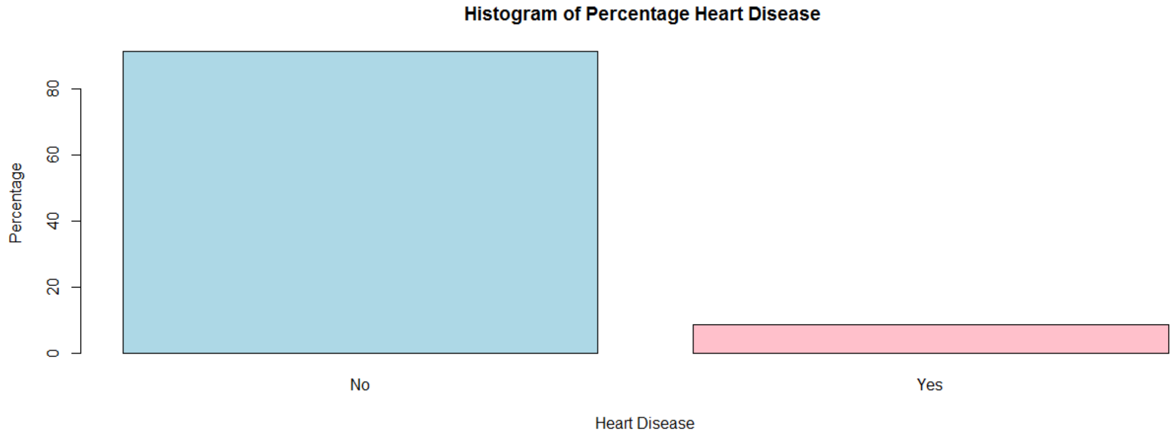


Figure 4: Histogram of the percentage heart disease

4.1. Evaluation Metrics

4.1.1. Confusion Matrix

Table 1: Confusion Matrix result summary

	Test	Train
GLM	Reference Prediction 0 1 0 57958 4870 1 526 604	Reference Prediction 0 1 0 231919 19503 1 2019 2396
Decision Tree	Reference Prediction No Yes No 54043 3579 Yes 4441 1895	Reference Prediction No Yes No 216004 14312 Yes 17934 7587
SVM	Reference Prediction No Yes No 41844 1033 Yes 16640 4441	Reference Prediction No Yes No 91583 21245 Yes 36074 106935

4.1.2. Accuracy

The table 2 shows the summary of accuracies of three methods (GLM, Decision Tree, and SVM) on both test and training datasets. GLM has the highest accuracy on both the test and train datasets, with values of 0.9153 and 0.9159, respectively, followed by Decision Tree with an accuracy of 0.8746 on the test dataset and 0.874 on the train dataset, and SVM with an accuracy of 0.7237 on the test dataset and 0.776 on the train dataset.

Table 2: Accuracy on Test and Train for three methods

	Test	Train
GLM	0.9153	0.9159
Decision Tree	0.8746	0.874
SVM	0.7237	0.776

4.1.3. AUC-ROC Curve

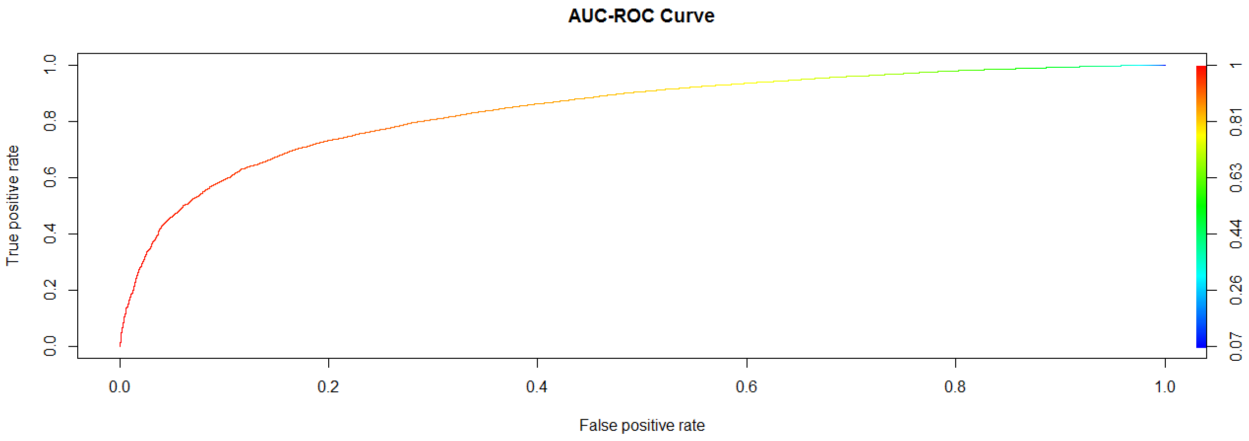


Figure 5: ROC Curve for GLM Method on Test Data

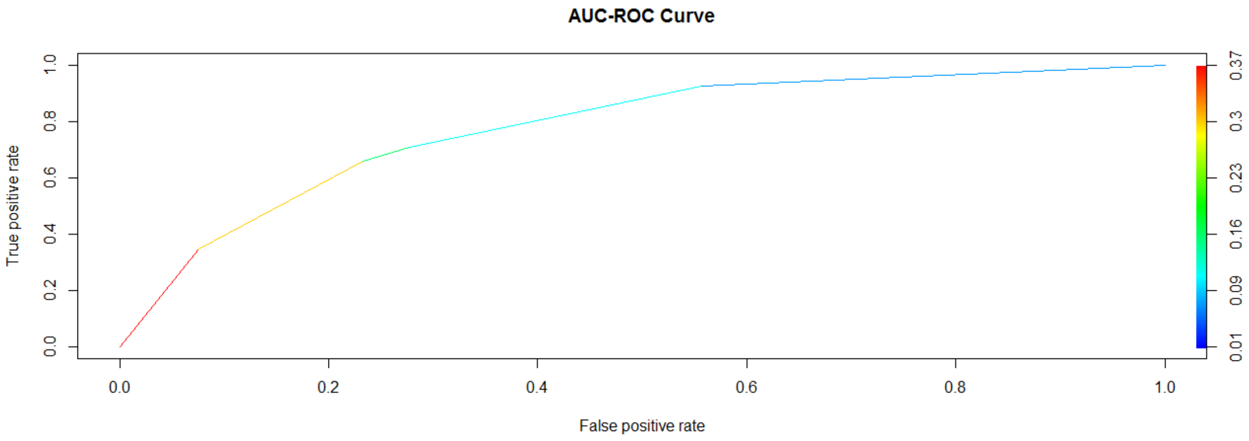


Figure 6: ROC Curve for Decision Tree Method on Test Data

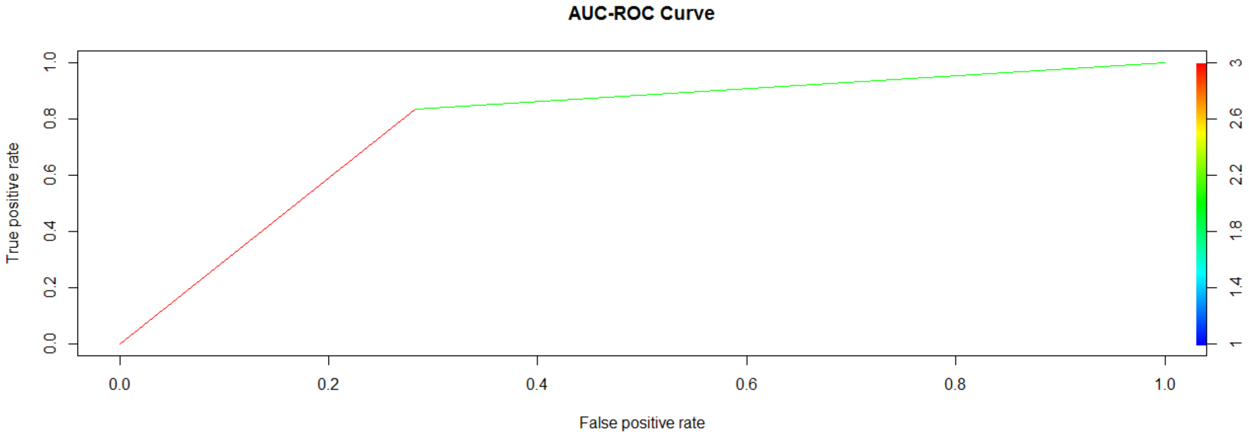


Figure 7: ROC Curve for SVM Method on Test Data

The table 3 displays the AUC-ROC (Area Under the Receiver Operating Characteristics) values for the test dataset for three methods (GLM, Decision Tree, and SVM). The GLM model has the highest AUC-ROC value with a value of 0.840, followed by the Decision Tree model with a value of 0.777 and the SVM model with a value of 0.763.

Table 3: AUC-ROC value on Test for three methods

Methods	Test
GLM	0.8402244
Decision Tree	0.7771811
SVM	0.7633837

The GLM model was found to have the best accuracy and ROC value among the three methods. One possible reason for the superior performance of the GLM model could be that it that are appropriate for the data being modeled. Additionally, the GLM model allows for the inclusion of only relevant predictors, which can result in a more robust model that focuses on the most important variables for predicting heart disease.

5. Conclusion

To conclude, the train accuracy for GLM, Decision Tree (DT) and SVM models were 0.916, 0.874 and 0.776 respectively. And the test accuracy for GLM, Decision Tree and SVM model were 0.915, 0.875 and 0.724 respectively. GLM had the best prediction with an accuracy of 0.916 and 0.915 in train and test data. The AUC-ROC for GLM was 0.84, for Decision Tree was 0.77 and lastly for SVM was 0.76. Again, it was observed the best AUC-ROC value was also for GLM. It is possible that the distribution of the dataset better bit the GLM model. Thus, GLM is better for predicting heart diseases.

References

1. Hoque E. Unit 3_ Classification [unpublished lecture notes]. DASC5420: Theoretical Machine Learning, Thompson Rivers University, lecture given 2023 Jan.
2. Hoque E. Unit 6--Support Vector Machines [unpublished lecture notes]. DASC5420: Theoretical Machine Learning, Thompson Rivers University, lecture given 2023 Feb.
3. Hoque E. Unit 11--Trees and Ensembles Methods [unpublished lecture notes]. DASC5420: Theoretical Machine Learning, Thompson Rivers University, lecture given 2023 Ap.
4. Pytlak K. Personal key indicators of heart disease [Internet]. Kaggle. 2022 [cited 2023Apr15]. Available from: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease