UTSA Journal of Undergraduate Research and Scholarly Works
Volume 8 **December 2022**

Performance of Machine Learning Algorithms for Heart Disease Prediction: Logistic Regressions Regularized by Elastic Net, SVM, Random Forests, and Neural Networks

Obehi Winnifred Ikpea and Dr. David Han
University of Texas at San Antonio, Texas
Department of Management Science and Statistics,

ABSTRACT

Heart disease, a medical condition caused by plaque buildup in the walls of the arteries, is the leading cause of death in the U.S. and worldwide. About 697,000 people suffer from this condition in the U.S. alone. This research project aims to assess and compare the performance of several classification algorithms for predicting heart disease so that the method can be considered as a clinical indicator of cardiovascular health. These methods include multiple logistic regression regularized with or without elastic nets, support vector machine, random forest, and artificial neural networks. A low prevalence of the disease is reflected in the data imbalance, and an oversampling technique is also suggested to deal with the computational challenges posed by this data imbalance.

***Keywords*:** Artificial neural networks, Elastic net, Heart disease prediction, Logistic regression, Machine learning algorithms, Random forests, Support vector machine

UTSA Journal of Undergraduate Research and Scholarly Works
Volume 8 December 2022

INTRODUCTION

Heart disease is the leading cause of death in the United States and globally [1]. About 17.9 million people die yearly from the disease and related conditions. It is an umbrella term for heart and blood vessel disorders such as coronary artery disease and other conditions. It is caused by plaque buildup in the walls of the arteries that supply blood to the heart and other parts of the body. The primary risk factors for heart disease include an unhealthy diet, physical inactivity, tobacco, excessive alcohol, and poor living conditions. In addition, these risk factors are present in individuals in the form of obesity, high glucose levels, and high blood pressure. The symptoms of heart disease vary amongst individuals and are not diagnosed early until the person experiences a heart attack, heart failure, or arrhythmia.

This study aims to investigate whether supervised learning algorithms can accurately predict heart disease occurrence given several risk factors. Supervised learning is a branch of statistical machine learning and artificial intelligence that trains algorithms on labeled input data in order to predict outcomes accurately. There are two types of supervised learning; regression and classification. This project focuses on the latter as it is desired to predict the case of heart disease in relation to some potential risk factors.

METHODOLOGY

Dataset Description The dataset used for this study is the stroke prediction dataset obtained from Kaggle, an online community for data collection, exploratory data analysis, and model building [3]. The stroke prediction dataset consists of 5,110 observations and 11 clinical features for predicting stroke events. Of these 11 features, we have two binary variables: hypertension and heart disease. For this study, the heart disease variable was the response variable, and 5 clinical features were the independent variables: 3 categorical variables (gender, marital record, residence type) and 2 continuous variables (average glucose level in blood and body mass index, BMI). The heart disease had 2 levels for classification; 0 for no heart disease and 1 for heart disease. All data preprocessing and analyses were done in the RStudio interface.

Dataset Preprocessing After selecting the required variables for the study, the dataset had to be preprocessed before conducting analyses. The preprocessed dataset consisted of 5,110 observations and 6 variables, of which 202 observations were missing (3.6% of the entire dataset). The missing values belonged to the BMI variable and they were replaced with its average values. The gender variable also had a class “Other,” which was removed from the dataset. After cleaning the data, the analyses began with 5,109 observations and 6 variables.

Train-Test Split The train-test technique is essential for evaluating the performance of any machine-learning algorithms. The heart disease dataset was split into the train and test datasets using a 70-30 ratio. So, 70% of the observations were randomly assigned to the training dataset and 30% to the test dataset. The training dataset had 3,576 observations while the test dataset had 1,533 observations. All machine learning algorithms were fitted on the training dataset.

RESULTS

UTSA Journal of Undergraduate Research and Scholarly Works
Volume 8 **December 2022**

Data Visualization A preliminary exploratory analysis of the training dataset was performed. The bar charts of binary predictors are given below, where blue colored bars indicate the cases of heart disease while red colored bars indicate no heart diseases. Figure 1 shows that there were more females than males in the dataset while the proportion of heart disease is smaller for females than for males. Figure 2 shows slightly more people living in the urban area, and Figure 3 shows the presence of relatively more married or previously married people in the dataset. Table 1 below shows a substantially large ratio of no heart disease cases to heart disease cases with the prevalence of 5.40%. This suggests that the dataset is severely imbalanced.

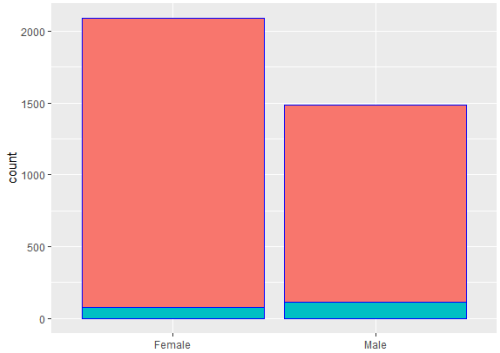

Figure 1. Gender distribution in the training dataset

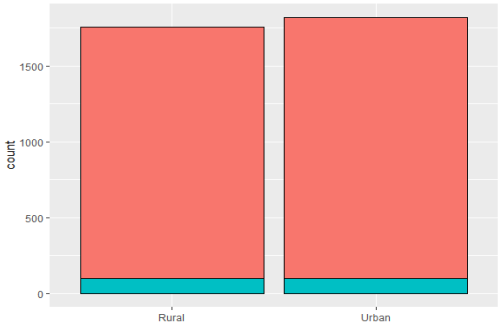

Figure 2. Residence type distribution in the training dataset

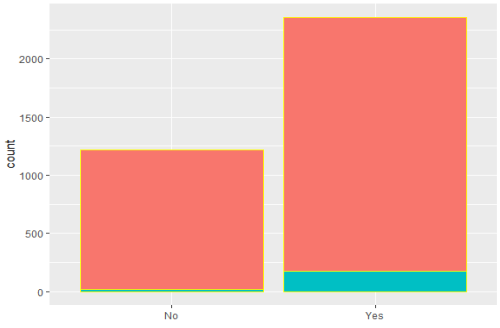

Figure 3. Marital record distribution (*i.e.*, ever married or not) in the training dataset

Table 1. Distribution of heart disease in the training dataset

Heart Disease	Frequency
Yes	198
No	3383

Oversampling with ROSE Due to the serious imbalance of the response variable, many model algorithms failed to converge or performed poorly on prediction of the heart disease. To handle the class imbalance of the training dataset, we tried different sample balancing techniques such as SMOTE (Synthetic Minority Oversampling Technique) and ROSE (Random Oversampling Examples). Unfortunately, SMOTE caused computational instability and so, we implemented ROSE, a bootstrap-based technique to create an artificially balanced binary class of heart disease in the training dataset. After oversampling, the heart disease variable had equal observations for both levels. The new training dataset eventually consisted of 6,766 observations, where the cases of heart disease and no heart disease had an equal frequency of 3,383 cases in each.

Model Training Four binary classification algorithms were fitted on the balanced training dataset. These algorithms include multiple logistic regression with or without elastic nets (including LASSO and ridge regression), support vector machine, random forest, and neural networks.

1) ***Multiple Logistic Regression Model*:**
A classic multiple logistic regression model was initially fitted with a full model composed of main effects and a cross-effect term for BMI and average glucose level. The backward elimination was performed to produce a final parsimonious model see Table 2. The final model fit only contained the main effect terms with AIC = 8101.1. Although the residence type was not found statistically significant, we decided to keep it in the model to be comprehensive. The ROC curve based on the logistic regression fit is displayed in Figure 4 below. The area under the curve (AUC) was 0.70, which suggests that there is a 70% chance the model will be able to distinguish between the heart disease and no-heart disease classes. It should be noted that the AUC value increased by using the ROSE technique for balancing the response variable.

Table 2. Estimate and standard error (s.e.) of the model coefficients with significance

Model Term	Estimate	s.e.	*z*-value	*p*-value
Intercept	-2.5287	0.1395	-18.131	$<10^{-15}$ **
BMI	-0.0092	0.0043	-2.170	0.030 *
Glucose Level	0.0108	0.0005	20.758	$<10^{-15}$ **
Gender (Male)	0.7738	0.0537	14.407	$<10^{-15}$ **
Marital Record (Yes)	1.4255	0.0701	20.327	$<10^{-15}$ **
Residence Type (Urban)	0.0270	0.0536	0.504	0.614

UTSA Journal of Undergraduate Research and Scholarly Works
Volume 8 December 2022

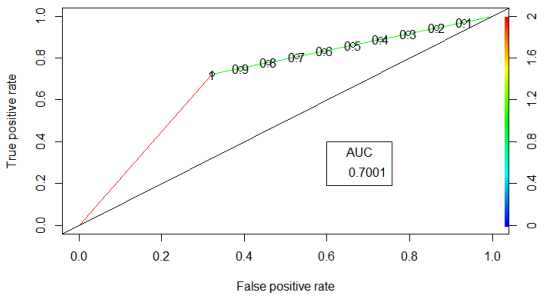

Figure 4. Receiver operating characteristic (ROC) curve of the logistic regression fit

2) ***Support Vector Machine*:**
Support vector machine (SVM), developed at AT&T Bell Laboratories, is another popular supervised learning model for robust classification and regression analysis. An SVM with linear Kernel was fitted for this classification task with 5 predictors. A 10-fold cross-validation was used, repeated 3 times. It should be noted that when the SVM was fitted to the original imbalanced dataset, the overall model accuracy increased but it was not able to predict any case of heart disease (*viz.*, sensitivity = 0%; positive predictive value, PPV = 0%), which is not useful for a prognostic medical purpose.

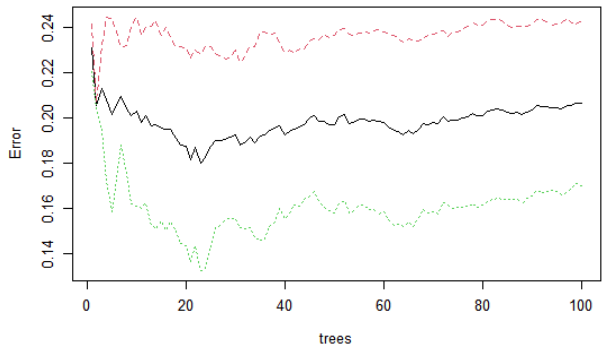

Figure 5. Training of a random forest

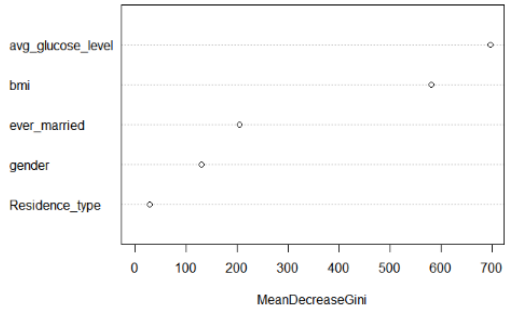

Figure 6. Variable importance plot for a random forest model

3) ***Random Forest Model*:**
Random forests (RF) or random decision forests is an ensemble learning method for classification and regression analysis. It operates by constructing a multitude of decision trees at training time, and for classification, the final output is determined by the class

selected by most decision trees. For the random forest model, we first created a plot to show the variable importance; see Figure 6 above. It is easy to see that the average glucose level is the most important variable in this case.

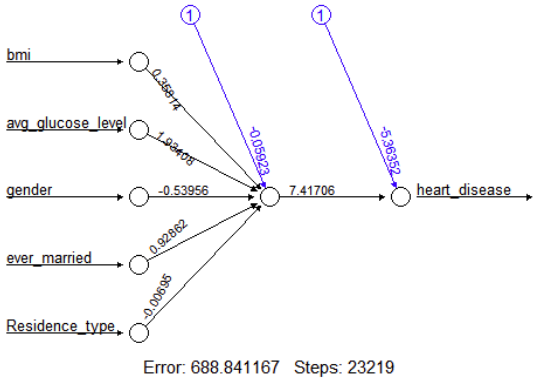

Figure 7. Artificial neural network model ANN(1) with a hidden layer and a single node

4) ***Artificial Neural Network Model*:**

Artificial neural networks (ANN), simply called neural networks or neural nets, are systemic models inspired by the biological neural networks. ANN is based on a collection of connected units or nodes called artificial neurons, organized in a hierarchy. It can be used for a variety of predictive tasks including regression and classification. Before fitting the neural network, we had to normalize the data in order to stabilize the gradient step to incorporate larger learning rates and a faster time for the model convergence [2]. First, we used the min-max normalization method on the numerical variables in the original imbalanced training dataset, and then we recoded the categorical variables into integers 0's and 1's. After the data transformation and normal-ization, the oversampling technique was performed on the training dataset. In order to manage the computation time, the first neural network model we fitted had one hidden layer with a single node in it;
see Figure 7 above.

Then, we increased the model complexity by constructing a neural network model with 2 hidden layers. The first hidden layer had 2 nodes while the second hidden layer had a single node; see Figure 8 below.

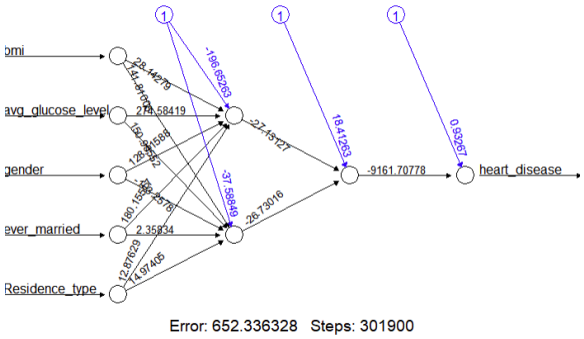

Figure 8. Artificial neural network model ANN(2,1) with 2 hidden layers; 2 nodes in the first layer and one node in the second layer

UTSA Journal of Undergraduate Research and Scholarly Works
Volume 8 **December 2022**

As the computation time was still reasonable, we further increased the model complexity by constructing a neural network with 2 hidden layers, each having 2 nodes; see Figure 9 below. About 6 hours were taken for the computation to converge and produce the final fit. For the reasons of computation time and potential overfitting, no further complex ANN (*i.e.*, deep learning) was considered in this work.

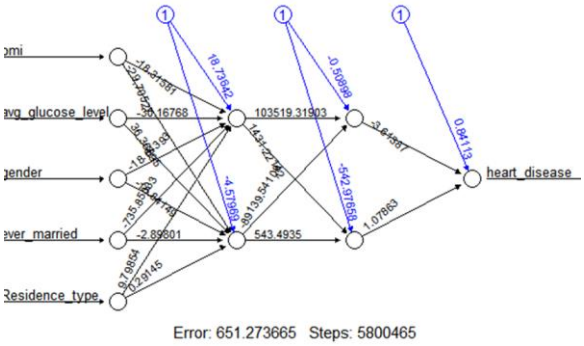

Figure 9. Artificial neural network model ANN(2,2) with 2 hidden layers, each having 2 nodes

Performance Assessment After fitting different statistical models and machine learning algorithms on the training dataset, it was realized that all the models achieved about the same accuracy rate of around 70% for predicting heart disease occurrence in the training data.

We then evaluated and compared the performance of all these machine learning algorithms for predicting heart disease occurrence in the imbalanced test data. Based on the confusion matrix table of each model, various metrics were calculated to assess the model performances, including the overall accuracy and its 95% confidence interval (CI), Cohen's κ (kappa), sensitivity (*a.k.a.* recall) and specificity, positive predictive value (PPV, *a.k.a.* precision), negative predictive value (NPV), and F1 score [4]. The results are summarized in Table 3 below. Using the formula $2 * \frac{Precision * Recall}{Precision + Recall}$, the F1 score over the test data was calculated to be approximately 0.20 for all the classifiers while Cohen's κ never achieved a value over 0.13, indicating a poor predictive power of very model.

Table 3. Summary of the predictive performance of various statistical models and machine learning algorithms over the test data

Model	Accuracy	95% CI		Cohen's κ	Sensitivity	Specificity	PPV	NPV
logistic	68%	(66%,	70%)	0.113	72%	68%	11%	98%
SVM	69%	(66%,	71%)	0.109	69%	69%	11%	98%
RF	75%	(73%,	77%)	0.125	59%	76%	12%	97%
ANN(1)	66%	(64%,	69%)	0.108	73%	66%	11%	98%
ANN(2,1)	70%	(68%,	72%)	0.118	69%	70%	12%	98%
ANN(2,2)	72%	(70%,	74%)	0.127	67%	72%	12%	97%

In an attempt to improve the predictive ability of a logistic regression model, we also fitted various regularized logistic regression models. This penalized model, known as the elastic net,

features a penalty term indexed by the value of α, $0 \leq \alpha \leq 1$. When $\alpha = 1$, the model becomes a LASSO regression with the L_1 penalty term, enabling variable selection and dimension reduction. When $\alpha = 0$, the model becomes a ridge regression with the L_2 penalty term. Various values of α were tried but the model performance did not change in a noticeable manner compared to the original logistic regression fit (*viz.*, no penalty); also see Table 2. This finding is evident from the fact that estimates of the model coefficients of these penalized logistic regression fits with elastic nets barely changed over a range of α; see Table 4 below.

DISCUSSION

Due the imbalanced nature of the test dataset with an overwhelmingly large number of class level 0 compared to that of class level 1, the original and regularized logistic regression models could not offer a good predictive power. This has been a well-known issue associated with severely imbalanced datasets as the model tends to predict the event of no primary interest (*e.g.*, no heart disease). As such a model is not clinically useful nor relevant for medical prognostic purposes, it necessitates balancing the dataset prior to analyses. For this study in particular, even after balancing the training data, all the resulting models produced the prediction accuracy between 68% and 75% over the test dataset, similar to the accuracy over the balanced training dataset; see Table 3. The estimates of PPV and NPV were almost identical across different models with poor predictability of the heart disease cases.

Nevertheless, it is interesting to note that the random forest model displayed the best accuracy over the test data although its recall is the worst among all the models we considered in this study. On the other hand, SVM performed only slightly better than the ordinary multiple logistic regression, regardless of the regularization.

The performance of the artificial neural network models is also noteworthy. As expected, the simplest ANN model with a single hidden node (*a.k.a.* perceptron) was the worst performing one in terms of the predictive accuracy. However, as the model structure accommodated more hidden layers and nodes, its performance over the normalized test data improved with the accuracy of 72%, closer to that of the random forest model. Therefore, including more layers and nodes could potentially improve the model accuracy even further although it may run into the issue of overfitting if not done carefully.

Table 4. Estimates of the coefficients of various logistic regression models regularized with elastic nets

Model Term	No Penalty	LASSO $\alpha = 1$	Elastic Net $\alpha = 0.8$	Elastic Net $\alpha = 0.6$	Elastic Net $\alpha = 0.4$	Elastic Net $\alpha = 0.2$	Ridge $\alpha = 0$
Intercept	-2.529	-2.517	-2.518	-2.515	-2.511	-2.504	-2.394
BMI	-0.009	-0.007	-0.008	-0.008	-0.008	-0.008	-0.006
Glucose Level	0.011	0.011	0.107	0.011	0.011	0.011	0.010
Gender (Male)	0.774	0.752	0.759	0.758	0.759	0.759	0.717

Marital Record (Yes)	1.426	1.387	1.399	1.397	1.397	1.395	1.305
Residence Type (Urban)	0.027	0.007	0.014	0.015	0.018	0.021	0.023

Another critical issue to address in practice of ANN is the computation time getting extensively and prohibitively longer with a larger and more complex ANN model, even if the model size is moderate. In such cases, the random forest model could be an attractive alternative with a reasonable computation time and comparative performance. This also implies that all the popularity and practical value of deep learning and artificial intelligence (A.I.) based on a neural network model could be unfounded without the possession of gigantic computational power and capacities.

CONCLUSION

In this study, we examined whether binary classification algorithms can predict heart disease occurrence using several risk factors. Upon handling the class imbalance with an oversampling method, 4 popular supervised learning algorithms were explored, including
multiple logistic regression regularized with or without elastic nets (including LASSO and ridge regression), support vector machine, random forest, and artificial neural networks. Similar but low F1 scores across the classifiers indicate that these binary classification algorithms are unsuitable for predicting the heart disease occurrences. This may be due to several reasons:

- The original dataset was designed for the stroke prediction but our study used it to predict the heart disease occurrence.
- The quality of the predictors was poor. It is believed that an insufficient number of independent variables and clinical features were provided for this study as well.
- The imputation method used to handle missing values in the dataset might be inadequate.
- The imbalanced nature of the dataset is a critical issue, and the sampling technique used to handle the problem is also imperfect.

For further investigation and analyses, we could consider the data augmentation method by including/merging clinically more relevant data. Along with different imputation techniques, use of a different undersampling or oversampling technique such as SMOTE, Tomek links, and weighted class techniques could be also explored to compare the results of prediction.

References

[1] Centers for Disease Control and Prevention (2021). "Cardiovascular diseases" (https://www.cdc.gov/globalhealth/healthprotection/ncd/cardiovascular-diseases.html).
[2] Cheng, Z.M. (2022). "Using normalization layers to improve deep learning models," *Machine Learning Mastery* (https://machinelearningmastery.com/using-normalization-layers-to-improve-deep-learning-models).
[3] Fedesoriano (2021). "Stroke prediction dataset," *Kaggle* (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset).
[4] Kamboj, S. (2020). "Performance metrics for evaluating a model on an imbalanced data set," *Medium* (https://medium.com/datasciencestory/performance-metrics-for-evaluating-a-model-on-an-imbalanced-data-set-1feeab6c36fe).