

A Comparison of some Discriminant Techniques in Predicting Blood Pressure

¹Yussuf T., ¹Adeleke B. L., ¹Oyeyemi G. M., ¹Adeleke M. O. and ²Kareem A. O.

¹Department of Statistics, University of Ilorin, Nigeria

²National Institute for Security Studies, Abuja, Nigeria

Corresponding Author: Dr. A.O. Kareem National Institute for Security Studies Abuja –
Nigeria. keemkareem@yahoo.com, Tel: +2347069719710

Abstract

The paper compared the performance of the Logistic Regression (LR), Fishers Discriminant Analysis (FDA) and the Support Vector Machines (SVM) in predicting high blood pressure. The variables used in the study are Age, Gender, BMI, Cholesterol and the smoking status. The SVM model was tuned to get the best parameter combination and cost function to avoid over fitting and under fitting of the model. The tuning model was used to compare the performance of the SVM model for sample sizes of 100, 500, 5000 and 7900. This approach was carried out for both the LR and FDA. The Data was divided into train and test data sets in the ratio 80:20 for all sample sizes considered to test the performance of the fitted models. The results from the sample sizes considered showed that for sample size of 100, the FDA performed better than the LR and SVM. But for sample sizes of 500, 5000 and 7900, the SVM performed better than both the LR and LDA. The area under the receiver operating curve showed 81.6% for the test data. This means that about 81.6% of the dataset was correctly predicted. The confusion matrix for the three approaches was computed. The result obtained showed the superiority of SVM to the other two methods.

Key words: Accuracy, Fisher Discriminant Analysis, Logistic Regression, Support Vector Machines, Performance

Introduction

Classification in statistics deals with predicting a predefined class membership accurately based on features of an individual. For instance, we may wish to classify and or predict the choice outcome of students on whether they will opt for a particular mode of transportation, or we may wish to diagnose a patient either healthy or diseased based on some measurements from the patient's tissue sample, or we may wish to classify a patients as hypertensive or not based on measurement on smoking habits, BMI etc. In general, goal of classification is to set a discriminating rule which is distinct enough to separate individual into one and only one of two or more groups [1].

One goal of classification is to learn the discrimination rule attaining the minimum error rate over novel cases. In the statistics literature, Fisher's linear discriminant analysis (FDA) and quadratic

discriminant analysis (QDA) are classical examples of a discriminant rule, and modern statistical tools include classification trees, logistic regression, neural networks, and kernel density based methods [2]. Logistic Regression (LR) appears as a robust alternative to FDA as it does not need any underlying assumption made on the distribution of the data. Hence, LR has always been suggested as the first choice to carry out data classification especially for a situation where the data is not normally distributed. Nevertheless, computing time of LR is much longer than the time taken by FDA, making it a less desirable alternative to FDA [3].

Support vector machines (SVM), on the other hand is a supervised learning technique for classification and regression purposes that has wide applications in data mining, engineering, and bioinformatics. The SVM works by constructing hyperplanes that separates cases of different class labels.



SVM can be used to solve non-linear regression problems, by the use of kernel functions; the kernel functions used includes linear, polynomial, radial basis function and sigmoid function. Since SVM works by pattern recognition and separating them into hyperplanes, it is quite tolerant of input data and also tolerant of the basic regression assumptions.

Logistic model is used as the based model because of its robustness and tolerance to underlying assumptions of normality; FDA on the other hand has been developed for normally distributed explanatory variables [4], while the SVM deals with pattern recognition and separating them into hyperplanes to predict new sets of data.

This paper therefore aims at comparing the performances of Fisher Discriminant Analysis, Logistic Regression Model and Support Vector Machine in predicting patients with high blood pressure using their predictive accuracy and receiver operating curve.

The paper is organized as follows: section two deals with material and methods where the three classification techniques' models were given as well as their advantages and limitations. The third section presents the data analysis and discussions of results while the fourth section gives the summary and conclusions.

Materials and methods

This section discusses the data used in the analysis and also examines the three techniques of classification in relation to their model specification, estimation and classification procedure. We also discuss the advantages and limitations of the methods

Data Description

The data used is secondary data which was culled via the internet from the home page of the centre for disease control and prevention, United States of America (USSAID) for the year 2012 [18]. It was originally collected by direct interview of some 14,030 adult patients under the national health and nutrition examination survey on hypertension, but due to missing observations in the data set and as a result of inconsistencies, the sample size was reduced to 7900. The sample was divided into 2 parts in the proportion 80:20 for the purpose of cross validation. The variables of concern in the data are age, sex, weight in pounds, height in inches, smoking status, cholesterol levels, average diastolic blood pressure and average systolic blood pressure.

The categorical variables namely; gender and smoking are coded as follows;

$$\begin{aligned} \text{Gender} &= \begin{cases} 0 & \text{for female} \\ 1 & \text{for male} \end{cases} \\ \text{Smoking} &= \begin{cases} 0 & \text{for smoking} \\ 1 & \text{for not smoking} \end{cases} \end{aligned}$$

The weight and height were used to calculate body mass index which is given as;

$$\text{BMI} = \frac{\text{weight in kilograms}}{(\text{height in metres})^2}$$

Presence of High Blood Pressure (HBP) is also coded as

$$\text{HBP} = \begin{cases} 0 & \text{for absence of HBP} \\ 1 & \text{for presence of HBP} \end{cases}$$

The Logistic Regression

The crucial limitation of linear regression is that it cannot deal with dependent variables that are dichotomous and categorical. Many interesting variables are dichotomous: for example, consumers make a decision to buy or not buy, a product may pass or fail quality control, there are good or poor credit risks, an employee may be promoted or not. A range of regression techniques have been developed for analyzing data with categorical dependent variables, including logistic regression and discriminant analysis. Logistic regression is regularly used rather than discriminant analysis when there are only two categories of the dependent variable. Discriminant analysis strictly requires the continuous independent variables however dummy variables can be used as in multiple regression [5 - 6].

Mainly, logistic regression however is appropriate for prediction of group membership, and it calculates the probability of success over the probability of failure and also provides knowledge of the relationships and strengths among two or more variables [7 - 8].

Model Fit and the Likelihood Function

Just as in linear regression, we are trying to find a best fitting line of sorts but, because the values of Y can only range between 0 and 1, we cannot use the least squares approach. The maximum likelihood is used instead to find the function that will maximize our ability to predict the probability of Y based on what we know about X. In other words, maximum likelihood finds the best values for formula. Likelihood just means probability. In logistic regression, two hypotheses are of interest: the null hypothesis, which is when all the coefficients in the regression equation take the value zero, and the alternate hypothesis that the model with predictors

currently under consideration is accurate and differs significantly from the null of zero. Log likelihood is the basis for tests of a logistic model. The likelihood ratio test is based on $-2\log$ likelihood ratio. It is a test of the significance of the difference between the likelihood ratio ($-2\log$ likelihood) for the researcher's model with predictors (called model chi square) minus the likelihood ratio for baseline model with only a constant in it.

Significance at the .05 level or lower means the researcher's model with the predictors is significantly different from the one with the constant only. It measures the improvement in fit that the explanatory variables make compared to the null model. Chi square is used to assess significance of this ratio. When probability fails to reach the 5% significance level, we retain the null hypothesis that knowing the independent variables or predictors has no increased effects in predicting the dependent. In the logistic regression model, the influence of X on Y is measured by the value of the slope of X which we have called β_1 . The null hypothesis, $H_0: \beta_1 = 0$ versus the alternative $H_1: \beta_1 \neq 0$ is of interest since if $\beta_1 = 0$, it implies X is not related to Y . Under the alternative $\beta_1 = \beta$, the logistic model becomes,

$$\log\left(\frac{P_1}{1-P_1}\right) = \beta_0 + \beta X \quad (1)$$

Under the null hypothesis, H_0 , this reduces to

$$\log\left(\frac{P_0}{1-P_0}\right) = \beta_0 \quad (2)$$

To test whether the slope is zero at a given value of X , the difference between these two quantities is formed giving

$$\beta_0 + \beta X - \beta_0 = \log\left(\frac{P_1}{1-P_1}\right) - \log\left(\frac{P_0}{1-P_0}\right)$$

Which reduces to

$$\begin{aligned} \beta X &= \log\left(\frac{P_1}{1-P_1}\right) - \log\left(\frac{P_0}{1-P_0}\right) \\ &= \log\left(\frac{P_1/(1-P_1)}{P_0/(1-P_0)}\right) \\ &= \log(\text{OR}), \text{ where OR is the odds ratio} \end{aligned} \quad (3)$$

This shows that the odds ratio (OR) of P_1 and P_0 is directly related to the slope of the logistic regression equation. It also shows that the value of the odds ratio depends on the value of X . For a given value of X , testing that B is zero is equivalent to testing OR is one. Since OR is commonly used and well understood, it is used as a measure of effect size in power analysis and sample size calculations.

The multiple logistic regression model relates the probability distribution of Y to two or more covariates $X_1, X_2, \dots, X_k$ by the formula

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (4)$$

Where P is the probability that $Y = 1$ given the values of the covariates. It is a simple extension of the simple logistic regression model that was just presented.

Fisher Discriminant Analysis

The Fisher's linear discriminant, is a method used in statistics, for pattern recognition and machine learning. It is used to find a linear combination of characteristics that characterizes or separates two or more classes of objects. Unlike principal component analysis which uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components, Fisher's discriminant analysis performs dimensionality reduction while preserving as much of the class discriminatory information as possible. A discriminant analysis is designed to enable the researcher to search for the linear equation that will maximize differences between the groups [9 -10] .

Recall the general form of a linear equation,

$$Y = a + bX,$$

and the form in linear multiple regression:

$$Y = a + b_1 X_1 + b_2 X_2 + \dots + b_p X_p \quad (5)$$

On the basis of the principle of least squares, the weights are selected so as to minimize the sum of squares errors. The linear equation referred to as the discriminant function takes the analogous form:

$$D = a + b_1 X_1 + b_2 X_2 + \dots + b_p X_p \quad (6)$$

Where D is the categorical variable to be predicted, specifically group membership. Where b_i denotes discriminant coefficients and X_i are predictor variables.

The objective of the use of the discriminant analysis is to form a linear equation for each group that maximizes the differences between the weighted group means which is also called group centroids. A discriminant analysis thus results in a discriminant function, or set of weights to be applied to the variables, in which the weights indicate the importance of each variable in contributing to group differences. The method also provides information regarding the statistical significance of the function as a whole and of the individual variable weights.

Application of Equation 6 to the calculation of the standardized discriminant score for the individual would result in equation 7 below;

$$D = a + b_1 X_{1i} + b_2 X_{2i} + \dots + b_p X_{pi} \quad (7)$$

In order to understand the concept of a centroid, consider the fact that the distribution of discriminant scores of members of a group, like any set of continuous scores, would have a certain distribution with a mean and standard deviation. The mean of discriminant scores within a group is known as the group centroid. The group centroid is calculated by applying the discriminant weights to the group

means on each variable. Equation 8 would be the formula for calculating the centroid of Group A:

$$\bar{D}_A = b_1 \bar{X}_{1A} + b_2 \bar{X}_{2A} + \dots + b_P \bar{X}_{PA} \quad (8)$$

In general, Discriminant analysis aims to identify characteristics which are important for differentiating units between groups and their classification accuracy. Each unit can be classified in only one category or group according to some predictor variables. The main theoretical assumptions for discriminant analysis are linearity and homoscedasticity [11].

The Probability that a particular case with discriminant score D belongs to the group i is calculated using Bayes theorem:

$$P(G_i|D) = \frac{P(D|G_i)P(G_i)}{\sum_{i=1}^k P(D|G_i)P(G_i)} \quad (9)$$

where (D/G_i) denotes conditional probability for D for a given group.

Support Vector Machines

Support Vector Machine (SVM) is a classification and regression method that combines computational algorithms with theoretical results; these two characteristics gave it good reputation and have promoted its use in different areas [12 -13]. In particular, SVM required less variables than LR to achieve an equivalent misclassification rate (MCR) [14] and [12],

Mathematical formulation of the SVM model

Assume training data $D = \{(x_i, y_i), i = 1, 2, \dots, N\}$ with $y_i \in \{-1, +1\}$ is separable by a hyperplane.

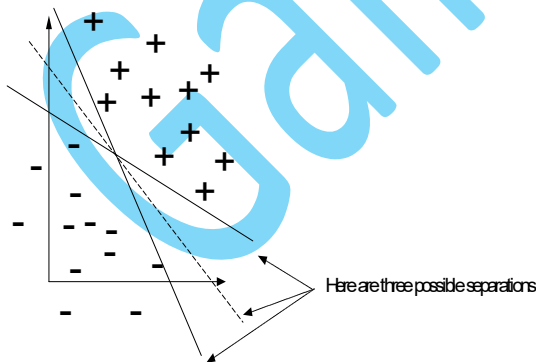


Figure 1: Hyperplane separating the data points

$$f(x) = w'x + b = w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_nx_n + b \quad (10)$$

While there can be an infinite number of hyperplanes that achieve 100% accuracy on training data, the question is what hyperplane is the optimal

with respect to the accuracy on test data? The approach is to increase the gap (margin) between positive and negative cases as much as possible.

Given $f(x)$, the classification is obtained as

$$\hat{y} = \text{sign}(f(x)) = \begin{cases} + & f(x) > 0 \\ -1 & f(x) < 0 \end{cases} \quad (11a)$$

w and b can result in the identical classification. For example, we can apply any scalar α such that:

$$\hat{y} = \text{sign}(\alpha(w'x + b)) = \text{sign}(w'x + b) \quad (11b)$$

Therefore there are many identical solutions.

Find $f(x) = (w'x + b)$ with maximal margin, such that for points closest to the separating hyperplane,

$$|w'x_i + b| = 1 \quad (\text{also called the support vectors})$$

and for other points, $|w'x_i + b| = 1 > 1$

Support Vector Machines: Learning Problem

Assuming a linearly separable dataset, the task of learning coefficients w and b of support vector machine $f(x) = (w'x + b)$ reduces to solving the following constrained optimization problem:

Find w and b that minimize: $\frac{1}{2} \|w\|^2$

subject to constraints: $y_i(w'x_i + b) \geq 1, \forall i$

This optimization problem can be solved by using the Lagrangian function defined as:

$$L(w, b, \alpha) = \frac{1}{2} w'w - \sum_{i=1}^N \alpha_i [y_i(w'x_i + b) - 1]$$

such that $\alpha_i \geq 0, \forall i$

Where $\alpha_1, \alpha_2, \dots, \alpha_N$ are Lagrange multipliers

The solution of the original constrained optimization problem is determined by the saddle point of $L(w, b, \alpha)$ which has to be minimized with respect to w and b and maximized with respect to α .

If $y_i(w'x_i + b) > 1$, the value of α_i that maximizes $L(w, b, \alpha)$ is $\alpha_i = 0$

If $y_i(w'x_i + b) < 1$ the value of α_i that maximizes $L(w, b, \alpha)$ is $\alpha_i = +\infty$. However, since w and b are trying to minimize $L(w, b, \alpha)$, they will be changed in such a way to make $y_i(w'x_i + b)$ at least equal to +1.

The conditions are as follow:

$$\alpha_i [y_i(w'x_i + b) - 1] = 0, \forall i \quad (11c)$$

The data points x_i with $\alpha_i > 0$ are called the support vectors.

The necessary conditions for the saddle point of $L(w, b, \alpha)$ are

$$\begin{aligned} \frac{\partial L}{\partial w_j} &= 0, \quad \forall j \\ \frac{\partial L}{\partial \alpha_i} &= 0, \quad \forall i \end{aligned} \quad (12)$$

or, alternatively, equation 12 can be expanded as equation 12a,

$$\nabla_w L = 0, \quad \nabla_\alpha L = 0 \quad (12a)$$

Solving for the necessary conditions results in

$$\begin{aligned} w &= \sum_{i=1}^N \alpha_i y_i x_i \\ \sum_{i=1}^N \alpha_i y_i &= 0 \end{aligned} \quad (12b)$$

By replacing $w = \sum_{i=1}^N \alpha_i y_i x_i$ into the Lagrangian function and by using $\sum_{i=1}^N \alpha_i y_i = 0$ as a new constrain the dual optimization problem can be constructed as

Find α that maximizes

$$\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j x_i' x_j \quad (13)$$

subject to $\sum_{i=1}^N \alpha_i y_i = 0, \alpha_i \geq 0, \forall i$

Equation 13 is a convex quadratic programming problem, so there is a global minimum.

Given the values $\alpha_1, \alpha_2, \dots, \alpha_N$ obtained, the final SVM predictor can be expressed from equation 12b as $f(x) = w'x_i + b = \sum_{i=1}^N \alpha_i y_i x_i' x_i + b$ (14) where

$$b = \frac{1}{|I_{\text{support}}|} \sum_{i \in I_{\text{support}}} \left(y_i - \sum_j \alpha_j y_j x_j' x_j \right)$$

And I_{support} is the set of support vectors.

To obtain the prediction, all data points from the training data are consulted and since $\alpha_i \neq 0$, only the support vectors are used in giving a prediction where $x_j' x_j$ is a scalar.

SVM solution for classification

Denote $\Phi: \mathcal{H}^M \rightarrow \mathcal{F}$ as a mapping from the original M-dimensional attribute space to the attribute space $\mathcal{F}$.

By solving the following dual problem

Find α that maximizes

$$\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \Phi(x_j)' \Phi(x_j) \quad (15a)$$

subject to

$$\sum_{i=1}^N \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C$$

the resulting SVM is of the form

$$\begin{aligned} f(x) &= w' \Phi(x_i) + b = \\ &= \sum_{i=1}^N \alpha_i y_i \Phi(x_i)' \Phi(x_i) + b \end{aligned} \quad (15b)$$

Table 1: Confusion matrix

		Target	
		Positive	Negative
Model	Positive	A	B
	Negative	C	D
		Sensitivity: $A/(A+C)$	Specificity: $D/(B+D)$
		+ Predictive Value: $A/(A+B)$ - Predictive Value: $D/(C+D)$	

The Positive Predictive Value (The true negative rate) is defined as the proportion of negatives cases that were classified correctly. It is the probability that the disease is present given a positive test result, Negative Predictive Value (The recall or true positive rate) is defined as the proportion of positive cases that were correctly identified. It is the probability that the disease is absent given a negative

test result. Although SVM are successful in dealing with highly dimensional attribute spaces, the fact that the SVM training scales linearly with the number of attributes, and considering limited memory space could largely limit the choice of mapping Φ [15].

The Kernel Trick

It allows computing scalar products of the form, $\Phi(x_i)' \Phi(x_i)$ in the original attribute space. The class of mappings Φ has the following property:

$$\Phi(x)' \Phi(y) = k(x, y), \quad (16a)$$

where K is a corresponding kernel function.

The dual problem finds α that maximizes

$$\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j k(x_i, x_j) \quad (16b)$$

subject to

$$\sum_{i=1}^N \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C$$

The resulting SVM is:

$$f(x) = w' \Phi(x_i) + b = \sum_{i=1}^N \alpha_i y_i k(x_i, x_j) + b \quad (17)$$

Model Evaluation and Validation

A diagnostic test is validated by comparing test results against a gold standard that establishes the true status of the subject. Validity is the ability of a test to indicate which individuals have the disease and which do not.

Test validation is an evaluation method used to determine the fitness of a test for a particular use and through it one can assess how good the test is at identifying subjects with and without a disease or condition. Validation involves calculating four objective measures of test performance, namely, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). The ideal diagnostic test would correctly identify subjects with and without the disease with 100% accuracy [16].

The table below gives the four objective measures of test performance

test result. Sensitivity is the ability of the test to identify correctly those who have a particular condition or disease. It is the proportion of true positives that are correctly identified by the test or the probability that an individual with the disease is correctly screened positive. Specificity is the ability of a test to correctly identify subjects without a particular condition or disease. It is the proportion of

true negatives that are correctly identified by the test. It is the probability that an individual without the disease is screened negative. The Accuracy is a qualitative term that describes how close the measurements are to the actual (true) value. It is the degree or percentage of resemblance to the true data set or phenomenon. It is the proportion of the total number of predictions that were correct

Receiver operating characteristic curve:

ROC curve graphs are useful tools for visualizing and evaluating classifiers. It is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold [17].

Data Analysis and Discussion of Results

This section presents results of all the analysis carried out in this study. In particular, Tables 1 and 2 give the summary of analyses on the data.

Table 4: Table of coefficients of the fitted logistic regression model

	Estimate	Std. Error	Pr(> z)	Odds ratios	Lower confidence interval	Upper confidence interval
Intercept	-7.3615	0.2997	<<0.01	0.000		
Age	0.0770	0.0025	<<0.01	1.079	1.074	1.084
Sex	0.1518	0.0743	< 0.05	0.866	0.760	0.986
BMI	0.0472	0.0064	<<0.01	1.043	1.032	1.055
Cholesterol	0.0041	0.0007	<<0.01	1.004	1.003	1.006
Smoking	-0.3141	0.0779	<<0.01	1.312	1.145	1.504

The results from Logistic regression show that males are more likely to have high blood pressure in comparison to females and the odds that a male will have high blood pressure increases by a factor of 0.866. Also non-smokers are less likely to have a high blood pressure than those who smoke. The odd that a person who does not smoke will have high blood pressure reduces by a factor of 1.312. And an increase in the body mass index and cholesterol makes an individual more likely to have high blood pressure with odds of 1.043 and 1.004 respectively. In summary, all the factors considered are found to significantly and good predicting factors for determining patient's blood pressure nature (Hypertensive or not hypertensive). Therefore people

Table 2: summary of Age, BMI and Cholesterol

	Age	BMI	Cholesterol
Minimum	20	13.29	59
Maximum	90	67.19	702

Table 3: summary of the categorical variables

Variables	Coding and summary	
HBP	0 = Not Hypertensive 6207	1 = Hypertensive 1693
GENDER	0 = Female 4727	1 = Male 3173
SMOKING	0 = Smoking 4021	1 = Not smoking 3879

Logistic Regression Model:

$$P(y = 1/X) = \frac{e^{(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5)}}{1 + e^{(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5)}}$$

$x_1 =$ Age; $x_2 =$ Sex; $x_3 =$ BMI; $x_4 =$ Cholesterol; $x_5 =$ Smoking

The training data set was used in training the model and the result of the model coefficients are given below

who are older, have higher BMI and cholesterol levels and smoke are more likely to be hypertensive.

The Confusion Matrix in evaluating the performance of the logistic regression model is shown in Table 5.

Table 5: Confusion matrix for the logistic regression

	0	1	Total
0	1223	28	1251
1	297	32	329
Total	1520	60	1580

Source: Author's computation 2018

From Table 5, the Positive Predictive Value is 98%. This means that 98% of those that were not having high blood pressure were correctly predicted

not to have High blood pressure by the model. While Negative Predictive Value is 9.7%, and that implies that 9.7% of those that have high blood pressure were correctly predicted to have high blood pressure by the model.

Also the Sensitivity is 80.5% which shows the model correctly identifies about 80.5% of the patients who have high blood pressure. And Specificity is 53.3%, that is about 53.3% cases of those who do have high blood pressure were screened as negative by the model.

The Correct classification (Accuracy) of logistic regression model is 79.4%, that is 1255 individual cases were correctly predicted/classified as hypertensive and not hypertensive. The percentage predictive accuracy is 79.4% which gives misclassification error as 20.6%.

ROC Curve

The closer the ROC curve is to the left hand corner of the graph the higher the overall accuracy of the test. From the figure above, the area under the ROC curve is 81.6% for the test data set and 82.4% for the training data set. This shows that the model predicted about 81.6% of the cases correctly.

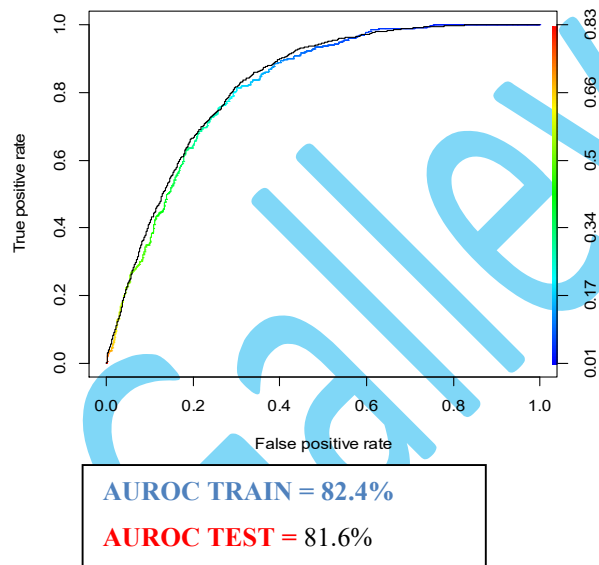


Figure 4.1: Receiver operating curve for the logistic regression

Fisher discriminant analysis:

For the whole data set, the confusion matrix and predictive accuracy of the Fisher linear discriminant model are as given in Table 6.

Table 6: Confusion matrix for the Fishers Linear Discriminant

	0	1	Total
0	1152	99	1251
1	227	102	329
Total	1379	201	1580

For Fisher discriminant model, the Positive Predictive Value as shown in Table 6 is 92.1%, this means that 92.1% of those that were not having high blood pressure were correctly predicted not to have High blood pressure in the model. The Negative Predictive Value is 31% which implies that 31.0 % of those that have high blood pressure were correctly predicted to have high blood pressure by the model.

Also the Sensitivity of the Fisher discriminant model is 83.5%, that is, the model correctly identifies about 83.5% of the patients who have high blood pressure. The Specificity of the model is obtained to be 50.7%, that is about 50.7% cases of those who do have high blood pressure were screened negative by the test.

The Accuracy of the model is 79.4% that is, 1254 individual cases were correctly predicted/classified as hypertensive and non-hypertensive. Since the percentage predictive accuracy is 79.4%, then the misclassification error is therefore 20.6%.

Support Vector Machine Analysis:

Firstly the model for SVM was tune using the tune function in the “e1071” library for the whole data set. Using 80% for training set and 20% for testing set. The best model was selected using the best.model function and the results are shown below; The best parameter combination was the eps-regression SVM-Type and SVM-Kernel of radial with a cost of 4 and epsilon 0.1

The correct classification rate of the model and confusion matrix is given in Table 7.

Table 7: Confusion matrix for the SVM, n = 7900, train set = 6320, test set = 1580

	0	1	Total
0	1209	42	1251
1	274	55	329
Total	1483	97	1580

From Table 4.7, the Positive Predictive Value of the SVM model is 96.6%, this implies that 96.6% of those that were not having high blood pressure were correctly predicted not to have High blood pressure in the model. While the Negative Predictive Value is obtained to be 16.7% which means, 16.7 % of those

that have high blood pressure were correctly predicted to have high blood pressure by the model.

Also, the Sensitivity of the SVM model is obtained to be 81.5%, that is, the model correctly identifies about 81.5% of the patients who have high blood pressure and the its Specificity is 56.7%, that is about 56.7% cases of those who do have high blood pressure were screened negative by the test.

The Accuracy of the model is 80.0%, that is, 1264 individual cases were correctly predicted/classified as been hypertensive or not hypertensive by the model. Since the percentage predictive accuracy is 80.0%, this implies that the misclassification error is 20.0%.

Comparison of LR, FDA and SVM for varying sample sizes

To determine the effect of sample size on the performances of the three classification methods, the percentage accuracy for correct prediction among the three methods for varying sample sizes, samples of $n = 100, 500, 5000$ and 7900 were compared. For each of the sample sizes, the data was divided into training and testing sets in the ratio 80% to 20% and the confusion matrix was computed for the testing data set for 200 iterations. The results of the simulation analysis are shown in Table 8 and the graphical presentation of the results is shown in Figure 2.

Table 8: Correct classification for LR, FDA and SVM

Method/sample size	N=100	N=500	N=5000	N=7900
LR	71.975%	76.65%	79.157%	79.4%
FDA	72.025%	76.905%	79.33%	79.4%
SVM	71.5%	77.235%	79.583%	80%

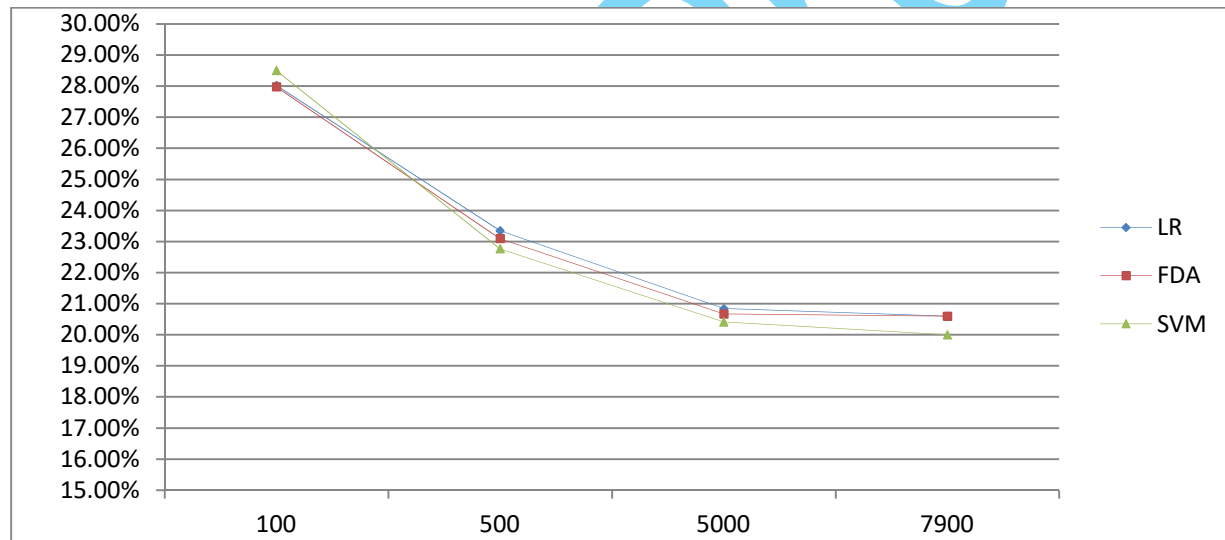


Figure 2: Plot of Misclassification Error Rate

Summary and Conclusion

This study considered Logistic regression, Fishers Discriminant Analysis and the Support vector machines to predicting the classes of individuals that have high blood pressure. The logistic regression was used to identifying the variables that contributes to the prediction of high blood pressure from the variables considered. From the logistic regression models, the major variables that contribute to predicting blood pressure models are age, gender, BMI, cholesterol and the smoking factor of the

individual. The odds ratio shows that males are more likely to have high blood pressure than females and also aging increases the chances of having high blood pressure. From the result of the analysis, a person who smokes has a higher chance of having high blood pressure than a person who does not and there is a increase in the odds that a person with higher BMI and Cholesterol will have high blood pressure.

In all approach considered in predicting individuals with high blood pressure the support vector machines is more efficient than the logistic regression and the

Fishers discriminant analysis. However as the sample sizes increases the three approach attain close values.

the support vector machines are more efficient in classifying and predicting the classes of high blood pressure. The SVM is able to classify correctly 80% of individuals as to having high blood pressure. Also the performance of the SVM was consistent as the sample sizes were increased. Although at sample size of 100 the FDA was seen to be better than both the LR and SVM.

References

- [1] Marcoulides, G.A. and Hershberger S.L. (1997): Multivariate Statistical Methods. New Jersey: CRS Press.
- [2] Betz, N. E. (1987): Use of Discriminant Analysis in Counseling Psychology Research: Journal of Counseling Psychology. 34(4), 393-403.
- [3] Oyeyemi, G. M., Oyebanji, L. A., Salawu, I. S. and Folorunsho, A. I. (2014): Comparison between Fisherian and Bayesian Approach to Classification Using two Groups. Scientia Africana. 13(1), 24 – 35.
- [4] Pohar, P. M., Blas, M. and Turk, S. (2004): Comparison of Logistic Regression and Linear discriminant Analysis: A simulation Study. Metodološki zvezki. (1), 143 – 161
- [5] Worth, A.P. and Cronin, M.T.D. (2003): The use of Discriminant Analysis, Logistic Regression and Classification Tree Analysis in the Development of Classification Models for Human Health Effects. Journal of Molecular Structure, Theochem. 622, 97-111
- [6] Oyeyemi, G. M. and Issa, S. (2018): Parametric and Pattern Recognition Models for Credit Scoring in Banking System. Journal of the Nigerian Association of Mathematical Physics 44, 163 – 174.
- [7] Dayton, M. C. (1992), “Logistic Regression Analysis”, University of Maryland Hawkins, D. M.(1982). Topics in Applied Multivariate Analysis. Cambridge university press, first edition New York
- [8] Held, E, Cape, J. and Tintle, N. (2016); Comparing Machine Learning and Logistic Regression Methods for Predicting Hypertension using a Combination of Gene Expression and Next Generation Sequencing Data; BMC Proc. 10(7), 10 – 34.
- [9] Hosmer, D.W. and Lemeshow, S. (1989): Applied Logistic Regression. New York: Wiley.
- [10] Neil, H.T. (2002). Applied Multivariate Analysis. Springer-verlag New York.
- [11] Oyeyemi, G. M. and Oyebanji, L. A. (2016): Classification into two groups with different cost of misclassification ratios. Journal of the Nigerian Association of Mathematical Physics. 35, 163 – 174.
- [12] Christofaro, D. G., Ritti-Dias, R. M., Fernandes, R. A., Polito, M. D., de Andrade, S.M., Cardoso, J. R., and de Oliveira, A. R. (2013); High Blood Pressure Detection in Adolescents by Clustering Overall and Abdominal Adiposity Markers; Universidade Estadual Paulista, Presidente Prudente, 96(6), 139 – 148
- [13] Khanna, D., Sahu, R., Baths, V. and Deshpande, B. (2015); Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease; International Journal of Machine Learning and Computing, 5(5), 215 – 229.
- [14] Verplancke, T., Van Looy, S., Benoit1, D., Vansteelandt, S., Depuydt, P., De Turck, F., and Decruyenaere, J. (2008): Support Vector Machine versus Logistic Regression Modeling for Prediction of Hospital Mortality in Critically Ill Patients with Haematological Malignancies. BMC Medical Information and Decision Making. 8, 125 – 138
- [15] Kanikar, P. and Shah, D. (2016): Prediction of Cardiovascular Diseases using Support Vector Machine and Bayesian Classification; International Journal of Computer Applications. 156(2), 8875 – 8887
- [16] Liong, C. Y. and Foo, S. F. (2014): Comparison of Linear Discriminant Analysis and Logistic Regression for Data Classification. AIP Conference Proceedings. 1522(1), 1159 – 1165.
- [17] Zweig, M. H. and Campbell; G. (1993): Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. American Association of Clinical Chemistry. 39(4), 221 – 233.
- [18] Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman. W. C., Green, L. A. and Izzo, J. L. (2003): The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. Hypertension. JAMA. 289(19):2560 - 2571.

Galley Proof