PROJECT

A STATISTICAL STUDY ON CLASSIFICATION TECHNIQUES

January 11, 2021

Benitta Susan Aniyan
Enrollment Number: EONFWL323564
Batch Number: 2020-4868

1.1 INTRODUCTION

Classification is a data mining function that assigns items in a collection to target categories or classes. The objective of classification is to accurately predict the target class for each case in the data. Classification models predict categorical class labels and prediction models predict continuous valued functions. The Classification technique is capable of processing a wider variety of data than regression and is growing in popularity.

1.2 MODELS

In my study, I used the following models: logistic regression and Receiver Operating Characteristic (ROC) Curve. These two are prediction models. For analysis, I'm using medical related data .The ground for this model choice is following: For most medical applications, the outcome is our interest is binary and the information can be expressed as probabilistic predictions. Predictions are hence absolute risks, which go beyond assessments of relative risks, such as regression coefficients, odds ratios or hazard ratios.

Logistic Regression, the most prevalent algorithm for solving industry scale problems, although it's losing ground to other techniques with progress in efficiency and implementation ease of other complex algorithms. The performance of prediction models can be assessed using a variety of different methods and metrics. Traditional measures for binary and survival outcomes include the Brier score to indicate overall model performance, the concordance (or c) statistic for discriminative ability (or area under the receiver operating characteristic (ROC) curve), and goodness-of-fit statistics for calibration.

1.3 DATA

The data consists of 3 data sets which took from UCI repository data set. The datasets are:

1. Dermatology data set – consists of 366 instances and 33 attributes.
2. Liver Disorders data set- consists of 345 instances and 7 attributes.
3. Abalone data set – consists of 4177 instances and 8 attributes.

1.4 OBJECTIVE OF THE STUDY

The objective is to compare the datasets with the two classification techniques and analyse which technique is best for each dataset.

2. METHODOLOGY

2.1. LOGISTIC REGRESSION

Logistic regression is an extension of simple linear regression. When the dependent variable is dichotomous or binary in nature, we can't use simple linear regression. Logistic regression is the statistical technique used to predict the relationship between predictors (our independent variables) and a predicted variable (the dependent variable) where the dependent variable is binary (e.g., sex [male vs. female], response [yes vs. no], score [high vs. low],etc…). There must be two or more independent variables, or predictors, for a logistic regression. The predictors, can be continuous (interval/ratio) or categorical (ordinal/nominal). All predictor variables are tested in one block to assess their predictive ability while controlling for the effects of other predictors in the model.

The logistic regression for Y as follows:

$$\text{Logit}(Y) = \ln\left(\frac{\pi}{1-\pi}\right)=\alpha+\beta_2X_2+\beta_2X_2$$

Where the Y intercept β's are regression coefficients, and X's are a set of predictors. α and β are typically estimated by the maximum likelihood (ML) method, which is preferred over the weighted least squares. The ML method is designed to maximize the likelihood of reproducing the data given the parameter estimates. Data are entered into the analysis as 0 or 1 coding for the dichotomous out-come, continuous values for continuous predictors, and dummy codings (e.g., 0 or 1) for categorical predictors.

2.1.1 CONFUSION MATRIX

In Machine Learning, the confusion matrix is also known as an error matrix. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa). It is a special kind of contingency table, with two dimensions i.e. actual and predicted, and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table).

2.1.2 TABLE OF CONFUSION

In predictive analysis, a table of confusion is a table with two rows and two columns. For a given classifier and a set of instances, there are four possible outcomes. That is, true positive, false negative, true negative, false positive.

If the instances is positive and it is classified as positive, it is counted as a true positive, if it is classified as negative, it is counted as a false negative. If the instance is negative and is classified as negative, it is counted as a true negative, if it is classified as positive, it is counted as a false positive. The numbers along the major diagonal represent the correct decisions made and the numbers of the other diagonal represent the errors.

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

Figure 1. Table of Confusion

True Positive Rate (TPR/ Recall/ Sensitivity) = $\frac{TP}{\sum Condition\ positive}$

False Positive Rate (FPR / Fall-out) = $\frac{FP}{\sum Condition\ negative}$

Accuracy= $\frac{TP+TN}{Total\ population}$

2.2 RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.

The ROC curve plots sensitivity against 1-specificity (probability of predicting an actual negative will be a positive).The best decision rule is high on sensitivity and low 1-specificity. It is a rule that predicts most true positives will be a positives and few true negatives will be a positive.

Sensitivity measures the proportion of actual positives that are correctly identified (e.g. the percentage of sick people who are correctly identified as having the condition).

Sensitivity= $\frac{TP}{\sum Condition\ positive}$

Specificity (True Negative Rate) measures the proportion of actual negatives that are correctly identified (e.g. the percentage of healthy people who are correctly identified as not having the condition).

Specificity=$\frac{TN}{\sum Condition\ condition}$

2.2.1 AREA UNDER THE CURVE (AUC)

AUC stands for "Area under the ROC Curve. That is, AUC measures the entire two-dimensional area underneath the entire ROC curve. AUC ranges in values from 0 to 1.A model whose predictions are 100% wrong has an AUC of 0.0 and predictions are 100% correct has an AUC of 1.0

3. DERMATOLOGY DATA SET

Dermatology is a branch of medicine that deals with the skin, hair, nails and its diseases.

3.1 ABSTRACT

The aim for this dataset is to determine the type of Erythemato -Squamous Disease.

There are six groups of Erythemato – Squamous diseases i.e. Psoriasis, Seboreic dermatitis, Lichen planus, Pityriasis rosea, chronic dermatitis and Pityriasis rubra pilaris.

Data Set Characteristics:	Multivariate	**Number of Instances:**	366	**Area:**	Life
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	33	**Date Donated**	1998-01-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	168521

3.2 DATA SET INFORMATION

This database contains 34 attributes, 33 of which are linear valued and one of them is nominal.

The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. The all six diseases share the clinical features of erythema and scaling. Usually a biopsy is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well. Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the 22 histopathological features evaluation. The values of the histopathological feature are determined by an analysis of the samples under a microscope.

In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the feature present in large amount, and 1, 2 indicate the relative intermediate values of the feature.

3.3 ATTRIBUTE INFORMATION

Clinical Attributes:

1: erythema

2: scaling

3: definite borders

4: itching

5: koebner phenomenon

6: polygonal papules

7: follicular papules

8: oral mucosal involvement

9: knee and elbow involvement

10: scalp involvement

11: family history (0 or 1) (nominal)

12: Age (linear)

Histopathological Attributes:

13: melanin incontinence

14: eosinophils in the infiltrate

15: PNL infiltrate

16: fibrosis of the papillary dermis

17: exocytosis

18: acanthosis

19: hyperkeratosis

20: parakeratosis

21: clubbing of the rete ridges

22: elongation of the rete ridges

23: thinning of the suprapapillary epidermis

24: spongiform pustule

25: munro microabcess

26: focal hypergranulosis

27: disappearance of the granular layer

28: vacuolisation and damage of basal layer

29: spongiosis

30: saw-tooth appearance of retes

31: follicular horn plug

32: perifollicular parakeratosis

33: inflammatory monoluclear inflitrate

34: band-like infiltrate

3.4 CODES

The analysis is done by using R, and the codes are given below:

```
ersd=read.csv("dermatology.csv")

ersd

attach(ersd)

head(ersd)

str(ersd)

logistic=glm(Family.history~.,ersd,family = 'binomial')

summary(logistic)

set.seed(0206)

library(caTools)

split=sample.split(ersd,SplitRatio = 0.75)

split

length(ersd$Family.history)

training=subset(ersd,split=="TRUE")

testing=subset(ersd,split=="FALSE")
```

```
dim(training)
logistic=glm(Family.history~.,training,family = 'binomial')
summary(logistic)
res=predict(logistic,testing,type='response')
table(Actualvalue = testing$Family.history,predictedvalue=res>0.5)
#accuracy--86.45%
res=predict(logistic,training,type='response')
library(ROCR)
ROCRPred=prediction(res,training$Family.history)
ROCRPref=performance(ROCRPred,'tpr','fpr')
plot(ROCRPref,colorize=TRUE,print.cutoff.at=seq(0.1,by=0.1),lwd=3,main='R
OC Curve')
t=.5
res_class=ifelse(res>t,1,0)
table(training$Family.history,res_class)
#sensitivity--97.46%
#specificity--27.27%
#accuracy--88.88%
abline(a=0,b=1)
auc=performance(ROCRPred,measure='auc')
auc <- auc@y.values[[1]]
auc
```

```
auc=round(auc,4)

auc

legend(.6,.2,auc,title="AUC")

#auc--.9092
```

3.5 OUTPUT AND INTERPRETATION

The table of confusion obtained by Logistic Regression and ROC curve is given in table 3.1

Table 3.1 Table of Confusion for dermatology data set

LOGISTIC REGRESSION

		PREDICTED VALUE	
		0	1
ACTUAL VALUE	0	81	2
	1	11	2

ROC

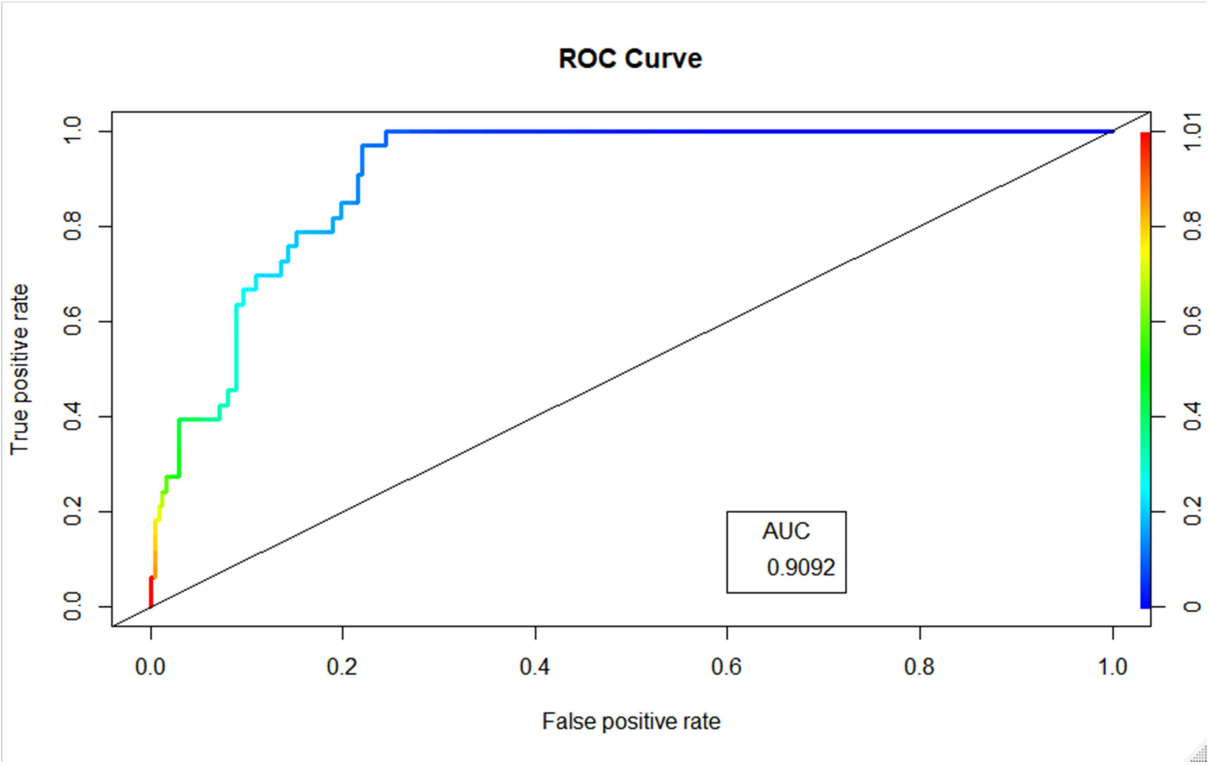
		PREDICTED VALUE	
		0	1
ACTUAL VALUE	0	231	6
	1	24	9

3.4.1. PARAMETERS VALUE

The accuracy of Logistic Regression and ROC is given in the table 3.2

Table 3.2 Parameters value for Dermatology dataset

DEPENDENT VARIABLE	TECHNIQUES	ACCURACY
FAMILY HISTORY	LOGISTIC REGRESSION	86.45%
	ROC	88.88%



3.5 RESULT

The Area under the ROC curve is .9092. And accuracy value of ROC=88.88%> accuracy of Logistic Regression=86.45%

So ROC curve is the best technique than logistic regression for dermatology dataset.

4. LIVER DISORDER DATA SET

4.1 ABSTRACT

To determine the presence or absence of liver disorders in a patient.

Data Set Characteristics:	Multivariate	**Number of Instances:**	345	**Area:**	Life
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	7	**Date Donated**	1990-05-15
Associated Tasks:	N/A	**Missing Values?**	No	**Number of Web Hits:**	142448

4.2 DATA SET DESCRIPTION

The first 5 variables are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each line in the dataset constitutes the record of a single male individual. The dataset does not contain any variable representing presence or absence of a liver disorder. Here, dichotomising the variable drinks number and used as a dependent variable for classification.

4.3 ATTRIBUTE INFORMATION

1. MCV- mean corpuscular volume

2. Alkphos alkaline phosphotase

3. sgpt alanine aminotransferase

4. sgot aspartate aminotransferase

5. gammagt gamma- glutamyl transpeptidase

6. Drinks number of half-pint equivalents of alcoholic beverages drunk per day

7. Selector field created by the BUPA researchers to split the data into train/test sets [47]

4.4 CODES

```
lids=read.csv("liver disorders1.csv")
```

```
lids

attach(lids)

head(lids)

str(lids)

logistic=glm(drinks.no.~.,lids,family = 'binomial')

summary(logistic)

set.seed(0206)

library(caTools)

split=sample.split(lids,SplitRatio = 0.75)

split

length(lids$drinks.no.)

training=subset(lids,split=="TRUE")

testing=subset(lids,split=="FALSE")

dim(training)

logistic=glm(drinks.no.~.,training,family = 'binomial')

summary(logistic)

res=predict(logistic,testing,type='response')

table(Actualvalue = testing$drinks.no.,predictedvalue=res>0.5)

#accuracy--78.78%

res=predict(logistic,training,type='response')

library(ROCR)

ROCRPred=prediction(res,training$drinks.no.)

ROCRPref=performance(ROCRPred,'tpr','fpr')

plot(ROCRPref,colorize=TRUE,print.cutoff.at=seq(0.1,by=0.1),lwd=3,main='ROC Curve')
```

```
t=.5

res_class=ifelse(res>t,1,0)

table(training$drinks.no.,res_class)

#accuracy--75.6%

abline(a=0,b=1)

auc=performance(ROCRPred,measure='auc')

auc <- auc@y.values[[1]]

auc

auc=round(auc,4)

auc

legend(.6,.2,auc,title="AUC")

#auc--.7635
```

4.5 OUTPUT AND INTERPRETATION

The table of confusion is obtained by the parameters, Logistic regression and ROC is given in the table 4.1

Table 4.1 Table of Confusion for Liver Disorder dataset

LOGISTIC REGRESSION

		PREDICTED VALUE	
		0	1
ACTUAL VALUE	0	67	2
	1	19	11

	PREDICTED VALUE

		0	1
ACTUAL VALUE	0	157	14
	1	46	29

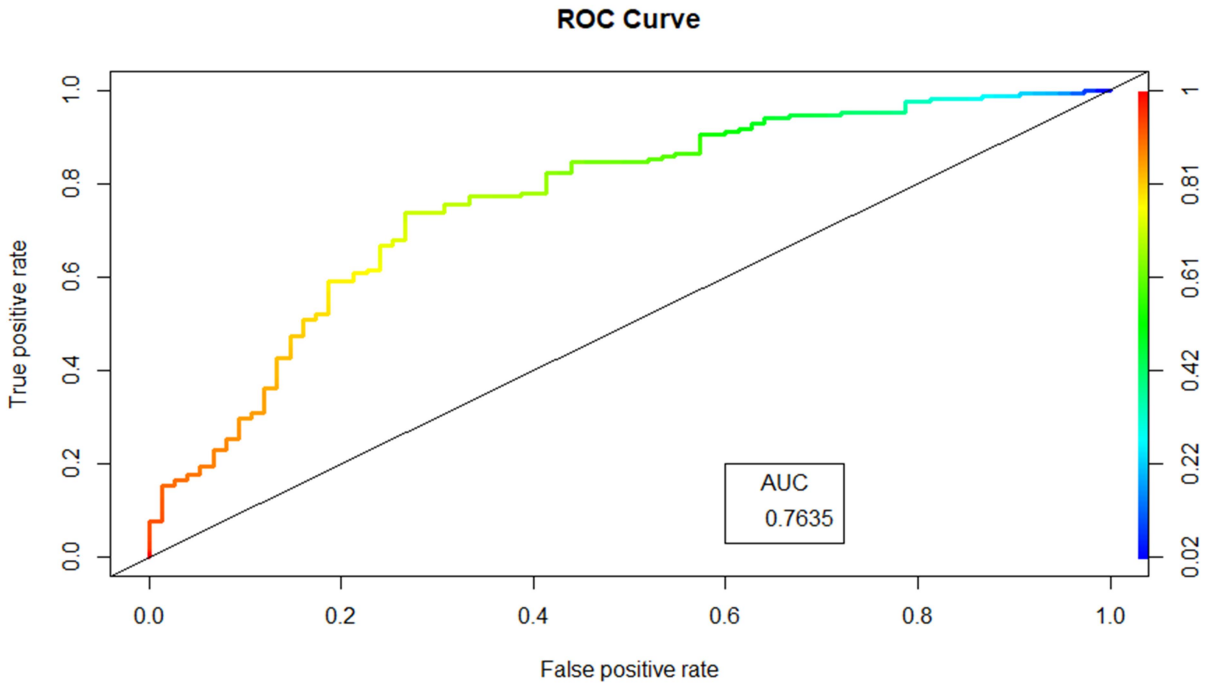
ROC

4.5.1 PARAMETERS VALUE

The accuracy of Logistic regression and ROC is given in the table 4.2

Table 4.2 Parameters value for Liver Disorder dataset

DEPENDENT VARIABLE	**TECHNIQUES**	**ACCURACY**
DRINKS NUMBER OF HALF PINT	LOGISTIC REGRESSION	78.78%
	ROC	75.6%



4.6. RESULT

The Area under the ROC curve is .7635.

Since, the accuracy of Logistic Regression=78.78%> accuracy of ROC curve=75.6%

Therefore, Logistic regression is the best technique for Liver Disorder dataset.

5. ABALONE DATA SET

5.1 ABSTRACT

Predict the age of abalone from physical measurements.

Data Set Characteristics:	Multivariate	**Number of Instances:**	4177	**Area:**	Life
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	8	**Date Donated**	1995-12-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	734835

5.2 DATA SET INFORMATION

Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope which is a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem. The number of rings is the value to predict.

5.3 ATTRIBUTE INFORMATION

Name / Data Type / Measurement Unit / Description

1. Sex / nominal / -- / M, F, and I (infant)
2. Length / continuous / mm / Longest shell measurement
3. Diameter/ continuous / mm / perpendicular to length
4. Height / continuous / mm / with meat in shell
5. Whole weight / continuous / grams / whole abalone
6. Shucked weight / continuous / grams / weight of meat
7. Viscera weight / continuous / grams / gut weight (after bleeding)
8. Shell weight / continuous / grams / after being dried
9. Rings / integer / -- / +1.5 gives the age in years

5.4 CODES

```
ablone=read.csv("abalone2.csv")
ablone
attach(ablone)
head(ablone)
str(ablone)
logistic=(Rings~.,ablone,family = 'binomial')
summary(logistic)
set.seed(0206)
library(caTools)
split=sample.split(ablone,SplitRatio = 0.75)
split
length(ablone$Rings)
training=subset(ablone,split=="TRUE")
testing=subset(ablone,split=="FALSE")
dim(training)
logistic=glm(Rings~.,training,family = 'binomial')
summary(logistic)
anova(logistic,test='Chisq')
res=predict(logistic,testing,type='response')
table(Actualvalue = testing$Rings,predictedvalue=res>0.5)
```

```
#acc--84.27
res=predict(logistic,training,type='response')
library(ROCR)
ROCRPred=prediction(res,training$Rings)
ROCRPref=performance(ROCRPred,'tpr','fpr')
plot(ROCRPref,colorize=TRUE,print.cutoff.at=seq(0.1,by=0.1),lwd=3,main='ROC
Curve')
t=.5
res_class=ifelse(res>t,1,0)
table(training$Rings,res_class)
#acc--84.23
abline(a=0,b=1)
auc=performance(ROCRPred,measure='auc')
auc <- auc@y.values[[1]]
auc
auc=round(auc,4)
auc
legend(.6,.2,auc,title="AUC")
#auc--.9117--<=8
```

5.5 OUTPUT AND INTERPRETATION

The table of confusion obtained by the parameters is given in table 7.1

Table 5.1 table of confusion for Abalone data

LOGISTIC REGRESSION

		PREDICTED VALUE	
		0	1
ACTUAL VALUE	0	352	136
	1	83	822

ROC

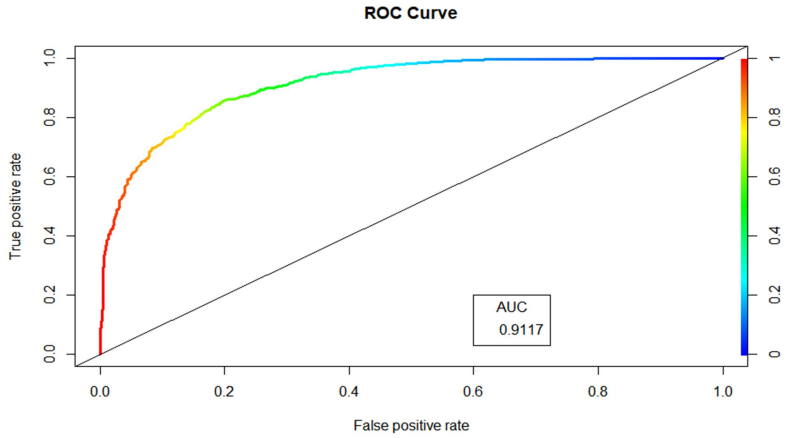
		PREDICTED VALUE	
		0	1
ACTUAL VALUE	0	668	251
	1	188	1677

5.5.1 PARAMETERS VALUE

The accuracy of Logistic Regression and ROC is given in table 7.2

Table 5.2 Parameters value for Abalone dataset

DEPENDENT VARIABLE	TECHNIQUES	ACCURACY
RINGS	LOGISTIC REGRESSION	84.27%
	ROC	84.23%



5.6 RESULT

The area under the curve is .9117. And the accuracy of Logistic Regression=84.27> Accuracy of ROC curve=84.23%

Logistic Regression has more accuracy than ROC curve.

So, Logistic Regression is the best technique for Abalone data set.

6. CONCLUSION

This study is about choosing the best classification technique between Logistic Regression and ROC Curve.
I have used the Dermatology dataset and applied two classification techniques that is, Logistic Regression and Receiver Operating Characteristic Curve. It has been interpreted as Receiver Operating Characteristic Curve performs better than Logistic Regression.

For the Liver Disorder dataset and applied two classification techniques. It has been interpreted as Logistic Regression performs better than Receiver Operating Characteristic curve.

I have used Abalone dataset and applied two classification techniques. It has been interpreted as Logistic Regression performs better than Receiver Operating Characteristic curve.

Overall, when we deals with classification techniques we can go with Logistic Regression, when we compare with Receiver Operating Characteristic curve. Since in most of the case, Logistic Regression gives best accuracy than Receiver Operating Characteristic (ROC) curve.