Case Study – Classification Problem

Submitted by: Saikat Ghosh

Due Date: 19th October 2015

Date of Submission: 16th November 2015

Supervisor's Remarks

Late Submission:

Plagiarism:

Completeness:

Quality of Content:

Results and Interpretations:

Additional Remarks:

What is Classification?

We want to classify an observation (univariate or multivariate) into one of several possible classes, or simply we want to estimate the probability given an observation that it belongs to one of the several possible classes.

Before we study classification, let us have an idea about the basic difference in classification and regression.

We know that Regression analysis is a statistical tool for the investigation and analysis of functional relationship between a set of independent variables and dependent variables. Regression typically assumes the Normality of observations.

In classification, we have response as **discrete** variable. Discrete variable can't be normally distributed the application of linear regression becomes invalid as the assumption of normality of the observations is no longer satisfied. Indeed the response follows a multinomial distribution. Moreover, the expectation of response becomes uninterpretable in terms of a linear function of the features and also the variances do not remain constant across observations and hence causing heteroscedasticity.

Thus such problems lie outside the domain of linear regression. Several tools namely naïve Bayes classifier, logistic classifier, discriminant analysis, nearest neighbor approach, neural network etc. are available to be deployed in such situations.

The classification problems are quite naturally divided into two types:

1. **Binary**, where response has two possible classes meaning by an observation either belongs to a class or it doesn't
2. **Multiclass**, where the response has more than two possible classes.

Binary Classification

Binary or **binomial classification** is the task of classifying the elements of a given set into two groups on the basis of a classification rule.

Some typical binary classification tasks are: Medical testing to determine if a patient has certain disease or not – the classification property is the presence of the disease; Quality control in factories; i.e. deciding if a new product is good enough to be sold, or if it should be discarded – the classification property is being good enough; information, namely deciding whether a page or an article should be in the result set of a search or not – the classification property is the relevance of the article, or the usefulness to the user.

Let us consider one such application of email spam filtering.

Email Spam Filtering

While new computer security threats may come and go, spam remains a constant nuisance for nonprofits. At a minimum, spam can interrupt your busy days, forcing you to spend time opening and deleting emails hawking herbal remedies or once-in-a-lifetime investment opportunities. In a more serious scenario, spam could unleash a nasty virus on your organization's network, crippling your servers and desktop machines.

The problem of classifying an email as either or non-spam (ham) is an interesting and relevant problem in the present world. To prevent email spam (a.k.a. unsolicited bulk email), both end users and administrators of email systems use various **anti-spam techniques**. Some of these techniques may be embedded in products, services and software to ease the burden on users and administrators. No technique is a complete solution to the spam problem, and each has trade-offs between incorrectly rejecting legitimate email (false positives) vs. not rejecting all spam (false negatives), and the associated costs in time and effort.

So to remedy this problem we use **Naïve Bayes Classifier.**

Naive Bayes spam filtering is a baseline technique for dealing with spam that can tailor itself to the email needs of individual users and give low false positive spam detection rates that are generally acceptable to users. Naive Bayes classifiers work by correlating the

use of tokens (typically words, or sometimes other things), with spam and non-spam e-mails and then using Bayesian inference to calculate a probability that an email is or is not spam.

An **advantage** of naive Bayes is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

Naïve Bayes Classifier in E-Mail Spam Filtering:

An email will have some information in itself about the fact that its spam or not. That information we have to find out and based on that we will estimate the probability of an email being spam. An email is nothing but a text, right?, which has words, symbols and numbers but all is text. This entire text including the no. of persons to whom it's been sent, the name of the persons to whom it's been sent, the cc list, the bcc list (which we can't see though), the subject line, the body message (header, main message, the signature and the postscripts) will have information about the message being spam or non-spam.
We can easily make a list of some common words which are generally seen in spam messages in a large frequency like congratulations, currency symbols, big numeric values, replica, derivative, claim, property, wealth etc. We have an idea of some common words. Or we can simply collect some 10-15 high frequency words from the spam emails in the training data.

This training is generally very large and maintains approximately a 50-50% ratio of the spam and non-spam message.

Now let's see how does the naïve Bayes classifier work? It's three step process viz.

- Computing the probability that the message is spam, knowing that a given word appears in this message;
- Computing the probability that the message is spam, taking into consideration all of its words (or a relevant subset of them);
- Dealing with rare words.

As a natural rule we do not consider the stop word like helping verbs, propositions etc. into consideration. For the purpose of above three steps we make use of the Bayes theorem in a absolute manner and hence the name Naïve Bayes Classifier.

Computing the probability that the message is spam, knowing that a given word appears in this message:

Let's suppose the suspected message contains the word "replica". Most people who are used to receiving email know that this message is likely to be spam, more precisely a proposal to sell counterfeit copies of well-known brands of watches. The spam detection software, however, does not "know" such facts; all it can do is compute probabilities.
The formula used by the software to determine that is derived from Bayes' theorem:

$$P(S|W) = \frac{P(W|S).P(S)}{P(W|S).P(S) + P(W|H).P(H)}$$

P($S|W$), is the probability that a message is a spam, knowing that the word "replica" is in it;
P(S), is the overall probability that any given message is spam;
P($W|S$), is the probability that the word "replica" appears in spam messages;
P(H), is the overall probability that any given message is not spam (is "ham");
P($W|H$), is the probability that the word "replica" appears in ham messages.

The 'spamicity' of a word
Recent statistics show that the current probability of any message being spam is 80%, at the very least which implies the following:

(S) = 0.80 and (H) = 0.2 : Prior Distribution/Probabilities

However, most Bayesian spam detection software makes the assumption that there is no *a priori* reason for any incoming message to be spam rather than ham, and considers both cases to have equal probabilities of 50%.

(S) = 0.50 and (H) = 0.50 : Prior Distribution/Probabilities

The filters that use this hypothesis are said to be "not biased", meaning that they have no prejudice regarding the incoming email. This assumption permits simplifying the general formula to the following:

$$P(S|W) = \frac{P(W|S)}{P(W|S) + P(W|H)}$$

This is functionally equivalent to asking, "what percentage of occurrences of the word "replica" appear in spam messages?". This quantity is called "spamicity" (or "spaminess") of the word "replica", and can be computed.

The number $(W|S)$ used in this formula is approximated to the frequency of messages containing "replica" in the messages identified as spam during the learning phase.
Similarly, $(W|H)$ is approximated to the frequency of messages containing "replica" in the messages identified as ham during the learning phase.

For these approximations to make sense, the set of learned messages needs to be big and representative enough. It is also advisable that the learned set of messages conforms to the 50% hypothesis about repartition between spam and ham, i.e. that the datasets of spam and ham are of same size. How this process does actually takes place? Suppose we have 500 spam emails and 500 ham emails (approximately half) we need to find $(W|S)$ then we can write it as follows:

$$P(W|S) = \frac{P(W \cap S)}{P(S)}$$

Where $(W \cap S)$ is calculated as the relative frequency of the word W in the spam messages. (S) is taken a priori. Similarly we can calculate $(H|S)$.

Of course, determining whether a message is spam or ham based only on the presence of the word "replica" is error-prone, which is why Bayesian spam software tries to consider several words and combine their spamicities to determine a message's overall probability of being spam.

Computing the probability that the message is spam, taking into consideration all of its words (or a relevant subset of them):

Most Bayesian spam filtering algorithms are based on formulas that are strictly valid (from a probabilistic standpoint) only if the words present in the message are independent events. This condition is not generally satisfied (for example, in natural languages like English the probability of finding an adjective is affected
by the probability of having a noun), but it is a useful idealization, especially since the statistical correlations between individual words are usually not known. On this basis, one can derive the following formula from Bayes' theorem:

$$p = \frac{p_1 p_2 \dots p_n}{p_1 p_2 \dots p_n + (1 - p_1)(1 - p_2) \dots (1 - p_n)}$$

$\boldsymbol{p}$ is the probability that the suspect message is spam;
$\boldsymbol{p_1}$ is the probability $(S|W_1)$ that it is a spam knowing it contains a first word (for example "replica");
$\boldsymbol{p_2}$ is the probability $(S|W_2)$ that it is a spam knowing it contains a second word (for example "watches");
$\boldsymbol{p_n}$ is the probability $(S|W_n)$ that it is a spam knowing it contains an nth word (for example "home").
This is the formula referenced by Paul Graham in his 2002 article. Spam filtering software based on this formula is sometimes referred to as a **Naïve Bayes classifier.**
The result p is typically compared to a given threshold; for our case taken to be 0.5 to decide whether the message is spam or not. If p is lower than the threshold, the message is considered as likely ham, otherwise it is considered as likely spam.

Case 1: Calculate the overall spamicity of the following emails and classify them as spam or non-spam.

Assume that spam and non-spam emails are equally probable in nature

Email 1:

Congratulations on winning the $ 100,000,000 in the lottery. To claim the prize send you contact details to lucky@xyz.com.

Email 2:

Everything is going fine. I will not be coming for summer holidays. Take care of yourself.

Step 1: To **compute the probability that the message is spam, knowing that a given word appears in this message**

Calculate the posterior conditional probabilities P(Spam| Word) and P(Ham|Word) from the given conditional probabilities P(Word|Spam) and P(Word|Ham) in the above emails using Baye's theorem.

The results are as follows:

Word	Email	P(Word | Spam)	P(Word | Ham)	Sum	pi	1-pi
Congratulations	1	0.8	0.2	1	0.8	0.2
winning	1	0.7	0.4	1.1	0.64	0.36
$	1	0.9	0.2	1.1	0.82	0.18
100000000	1	0.7	0.1	0.8	0.88	0.13
lottery	1	0.6	0.2	0.8	0.75	0.25
claim	1	0.6	0.3	0.9	0.67	0.33
prize	1	0.6	0.4	1	0.6	0.4
send	1	0.5	0.5	1	0.5	0.5
you	1	0.7	0.3	1	0.7	0.3
contact	1	0.5	0.5	1	0.5	0.5
details	1	0.6	0.6	1.2	0.5	0.5
Everything	2	0.2	0.7	0.9	0.22	0.78
going	2	0.2	0.7	0.9	0.22	0.78
fine	2	0.7	0.5	1.2	0.58	0.42
I	2	0.5	0.5	1	0.5	0.5
coming	2	0.5	0.5	1	0.5	0.5
summer	2	0.6	0.6	1.2	0.5	0.5
holidays	2	0.8	0.4	1.2	0.67	0.33
Take	2	0.7	0.6	1.3	0.54	0.46
care	2	0.2	0.2	0.4	0.5	0.5
yourself	2	0.8	0.7	1.5	0.53	0.47

Conclusion

Computing the overall spamicity using the formula specified above for both the emails, we get :

* For **Email 1**, **p = 0.999784** implying that Email 1 can be regarded as spam
* For **Email 2, p = 0.233577** implying that Email 2 cannot be regarded as spam

Logistic Regression

Consider the multiple linear regression model (MLRM) for sample observations.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i \; ; \; i = 1, \ldots, n$$

Equivalently

$$Y_i = \boldsymbol{X_i^T \beta} + \epsilon_i \; ; \; i = 1, \dots, n$$

where $\boldsymbol{X_i} = (X_{1i}, \dots, X_{pi})^T$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$.

Suppose Yi given $X_1, \dots, X_p$ is a categorical random variable. For simplicity let $Yi \mid X1, \dots, X_p$ is an indicator variable defined as follows.

$$Y_i \mid X_1, \dots, X_p = \begin{cases} 1 & \text{if } i\text{th sample observation possesses some attribute under study} \\ 0 & \text{otherwise} \end{cases}$$

Clearly $Yi \mid X_1, \dots, X_p \sim \text{Bin}(1, \pi i)$ where $\pi i = P(Yi = 1 \mid X1, \dots, Xp)$ and $0 \leq E(Yi \mid X1, \dots, Xp) = \pi i \leq 1$.

We model the conditional probabilities using a non-linear function of the independent variables of the following form.

$$P\left(Y_i = 1 \middle| X_1, \dots X_p\right) = g\left(\boldsymbol{X_i^T \beta}\right)$$

The function g(.) is called as a link function. The most common link function is the **logit link function or logistic function or sigmoid function** which is defined as follows.

$$g(z) = \frac{\exp(z)}{1 + \exp(z)} = \frac{1}{1 + \exp(-z)}$$

In terms of the logit link function the model under study is actually given by

$$Y_i = \frac{\exp[\boldsymbol{X_i^T \beta}]}{1 + \exp[\boldsymbol{X_i^T \beta}]} + \epsilon_i \quad .$$

After specifying an appropriate model for modeling a binary response variable, the next task is to estimate the parameters.

The maximum likelihood estimator of $\boldsymbol{\beta}$ is defined as maxima of $[\; \mid Yi \; ; \; i = 1, \dots, n \;]$ with respect to $\boldsymbol{\beta}$, i.e.

$$\text{MLE}\,[\boldsymbol{\beta}] = \underset{\boldsymbol{\beta}}{\text{argmax}}\, L\left[\, \boldsymbol{\beta} \;\middle|\; Y_i \; ; \; i = 1, \dots, n \,\right] = \hat{\boldsymbol{\beta}}$$

Explicit maximization of likelihood function is not possible, although iterative methods like Iterative Reweighted Least Squares (IRLS) can be used.

Hence using invariance property of MLE

$$\text{MLE}\,[\pi_i] = \frac{\exp[\boldsymbol{X_i^T \hat{\beta}}]}{1 + \exp[\boldsymbol{X_i^T \hat{\beta}}]} = \hat{\pi}_i$$

The estimated conditional probability of $Yi = 1$ is given by $P\hat{}(Yi = 1 \mid X1, \dots, Xp) = \hat{\pi i}$. Hence a classification rule can be given as follows:

$$\hat{\pi}_i \geq 0.5 \text{ then } \hat{Y}_i = 1 \text{ otherwise } \hat{Y}_i = 0$$

Checking Adequacy of Logistic Regression Model

A simple 2 × 2 classification table termed as **confusion matrix** can be constructed.

Confusion Matrix		Estimated Response $\hat{Y}$	
		0	1
Observed Response Y	0	f_{00}	f_{01}
	1	f_{10}	f_{11}

Following simple measures of goodness can be defined

1. Percentage of Misclassification = $\frac{f_{01}+f_{10}}{n} \times 100$:
2. Sensitivity = $\frac{f_{11}}{f_{11}+f_{01}}$:
3. Specificity= $\frac{f_{00}}{f_{00}+f_{10}}$

Odds Ratio

Odds Ratio (OR) is a measure of association between the two attributes *A* and *B*, say defined as follows.

$$OR = \frac{\text{Odds in favor of } A \text{ in the presence of B}}{\text{Odds in favor of } A \text{ in the absence of B}}$$

In case of binary regression odds ratio is given by;

$$OR = \frac{\text{Odds in favor of } Y_i = 1 \text{ at } X_i = 1}{\text{Odds in favor of } Y_i = 1 \text{ at } X_i = 0} = \exp[\beta_1]$$

Hosmer–Lemeshow test

The Hosmer-Lemeshow test is a statistical test for goodness of fit for the logistic regression model. The data are divided into approximately ten groups defined by increasing order of estimated risk. The observed and expected number of cases in each group is calculated and a Chi-squared statistic is calculated as follows:

$$\chi_{HL} = \sum_{g=1}^{n} \frac{(O_g - E_g)^2}{E_g(1 - E_g/n_g)}$$

with O_g, E_g and n_g the observed events, expected events and number of observations for the g^{th} risk decile group, and n the number of groups. The test statistic follows a Chi-squared distribution with$n-2$ degrees of freedom.

A large value of Chi-squared (with small p-value < 0.05) indicates poor fit and small Chi-squared values (with larger p-value closer to 1) indicate a good logistic regression model fit.

The **Contingency Table for Hosmer and Lemeshow Test** table shows the details of the test with observed and expected number of cases in each group.

Following are some common applications where logistic regression is used as a classification technique:

(i) Spam Detection: Spam/Non-Spam
(ii) Tumor Examination: Malignant/Benign
(iii) Online Transaction: Fraudulent/Non-Fraudulent
(iv) Email Labeling: Work, Friends, Family etc.

Case 2: Prediction of Cancer due to Smoking using Logistic Regression

Given the data on a binary response variable telling us whether the cancer is present or not and a single binary independent variable telling whether the person smokes or not we want to predict the possibility of cancer due to smoking. In nutshell we want to know "how more likely is a person to have cancer if he/she smokes rather he/she doesn't". We are supposed to do the following:

A study was performed on lung cancer possibility due to smoking habits. Data on presence/absence of two attributes viz. lung cancer and smoking was collected for 24 individuals.

Lung Cancer (Y)	Smoking (X)
1	0
0	0
1	1
0	1
0	0
1	0
1	0
0	0
0	1
1	1
0	0
0	0
1	1
0	1
1	1
1	0
1	1
0	1
1	1
1	0
1	1
1	1
1	0
1	1

Step 1: Build a logistic regression model for cancer possibility using smoking as an independent variable

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	**Smoking**	.770	.823	.875	1	.350	2.160
	Constant	-.182	.606	.091	1	.763	.833

a. Variable(s) entered on step 1: Smoking.

The fitted model is: $\hat{\pi}_i = \frac{e^{-0.182+0.770x}}{1+e^{-0.182+0.770x}}$

Where,

Y_i is the dependent variable, X is the independent variable

Step 2: Test for the Significance individual independent variables

The p value for both the parameters viz. β_o and β_1 is 0.763 and .350 respectively, that is greater than 0.05. Hence, we conclude that both the parameters are insignificant and thus X is not significantly affecting the response.

Step 3: Construct the Confusion (Classification) Table and report the percentage of correct classification in the given emails. Also calculate specificity and sensitivity of the model.

Classification Table[a]

		Predicted		
		Cancer		
Observed		**0**	**1**	**Percentage Correct**
Cancer	**0**	6	5	54.5
	1	5	9	64.3
Overall Percentage				60.0

a. The cut value is .500

From the classification table, we compute the following:

- Percentage of correct classification = 60.0% which shows that the model classifies the predicted values correctly for more than half of the times.
- Sensitivity = 0.64 showing that proportion of cases correctly identified by the test as meeting the criterion of smoking , is not very high.
- Specificity = 0.54 showing that proportion of cases correctly identified by the test as not meeting the criterion of smoking, is not very high.

Step 4: For each person obtain the probability of him/her having cancer and hence the prediction of cancer using the Logistic Classifier you have built

Probability that an individual is a smoker is given by:

$$\hat{\pi}_i = \frac{e^{-0.182+0.770x}}{1+e^{-0.182+0.770x}}$$

= 0.64286

The probability of each person having cancer and hence the prediction of cancer using the Logistic Classifier is given in the following table:

Case Summaries[a]

		Predicted probability	Predicted group
1		.45455	0
2		.45455	0
3		.64286	1
4		.64286	1
5		.45455	0
6		.45455	0
7		.45455	0
8		.45455	0
9		.64286	1
10		.64286	1
11		.45455	0
12		.45455	0
13		.64286	1
14		.64286	1
15		.64286	1
16		.45455	0
17		.64286	1
18		.64286	1
19		.64286	1
20		.45455	0
21		.64286	1
22		.64286	1
23		.45455	0
24		.64286	1
25		.64286	1
Total	N	25	25

a. Limited to first 100 cases.

Step 5: Estimate the odds ratio and interpret it.

Odds Ratio for the model: $\exp(\beta_1) = 2.160$

Since **OR > 1**, exposure is associated with higher odds of outcome i.e. if a person smokes, he/she is more likely to have lung cancer than the person who does not smoke.

Case 3: Skull Type Prediction using Logistic Regression

We are interested in predicting the type of skull of humans as one of two possible types I and II based on some five physical measures available related to the skulls.

Step 1: Build a logistic regression model for classifying a human skull as Type I/Type II using the given independent variables.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	**X1**	-.008	.018	.185	1	.668	.992
	X2	-.047	.033	2.039	1	.153	.954
	X3	-.007	.020	.113	1	.737	.993
	X4	-.006	.024	.054	1	.816	.994
	X5	.022	.020	1.245	1	.264	1.022
	Constant	1.149	2.826	.165	1	.684	3.155

a. Variable(s) entered on step 1: X1, X2, X3, X4, X5.

: *The fitted model is:* $\hat{\pi}_i = \frac{e^{1.149-0.008x1-0.047x2-0.007x3-0.006x4+0.22x5}}{1+e^{1.149-0.008x1-0.047x2-0.007x3-0.006x4+0.22x5}}$

We conclude that p-values of all the five variables i.e.$x_1,\ x_2, x_3,\ x_4,\ x_5$ are greater than 0.05 therefore all the independent variable are not significant at 5% level of significance.

Step 2: Test for the Significance individual independent variables

From the above table we can see that the p-values of all the variables X_1,X_2,X_3,X_4 and X_5 are greater than 0.05,so we may accept our H_0.
Thus none of the independent variables is significantly affecting the response.

Step 3: Test for the overall Logistic Regression using Hosmer and Lemeshow Test

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.601	8	.294

We conclude that p-value is (0.294) > 0.05, so Logistic Regression model gives a good fit to the given data.

Contingency Table for Hosmer and Lemeshow Test

		Skull = 0		Skull = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	2	1.841	0	.159	2
	2	2	1.774	0	.226	2
	3	2	1.446	0	.554	2
	4	0	1.374	2	.626	2
	5	2	1.226	0	.774	2
	6	1	1.073	1	.927	2
	7	0	.891	2	1.109	2
	8	1	.778	1	1.222	2
	9	1	.451	1	1.549	2
	10	0	.146	1	.854	1

Classification Table[a]

		Predicted		
		Skull		
Observed		0	1	Percentage Correct
Skull	0	9	2	81.8
	1	3	5	62.5
Overall Percentage				73.7

a. The cut value is .500

Percentage of correct classification: 0.736

Sensitivity of the model: 0.62 which implies that 62 out of 100 persons were predicted to have Skull Type II

Specificity of the model:0.81 which implies that 81 out of 100 persons were predicted to have same Skull type I.

Probability of skull being Type I or Type II, and hence predict the skull Type using the Logistic Classifier

Case Summaries[a]

	Predicted probability	Predicted group
1	.58476	1
2	.82867	1
3	.07618	0
4	.72021	1
5	.08238	0

6		.25299	0
7		.32244	0
8		.52374	1
9		.30390	0
10		.34643	0
11		.60310	1
12		.47311	0
13		.61877	1
14		.08545	0
15		.30121	0
16		.85355	1
17		.45430	0
18		.42803	0
19		.14077	0
Total	N	19	19

a. Limited to first 100 cases.

For a set of five physical measures given for a new skull in the dataset **Skull Type Prediction,** predict the skull type using the Logistic Classifier we have built:

Skull Type:	X_1	X_2	X_3	X_4	X_5
	171	134	130	69	130

Estimated coefficients:	β_0	β_1	β_2	β_3	β_4	β_5
	1.149	-.008	-.047	-.007	-.006	.022

Substituting the given values of independent variables in the given equation, we get:

$$\hat{\pi}_i = \frac{\exp[X_i^T \hat{\beta}]}{1+\exp[X_i^T \hat{\beta}]}$$

Now, $\hat{\pi}_i = 0.007 < 0.5$, therefore we would take *Skull Type I*

Sentiment Analysis using Logistic Regression – What makes a US presidential Candidate Win?

We are interested here in knowing that depending upon what and how a politician give speeches, his/her chances of winning the elections are affected. The idea here is similar to the email spam detection. The speech and more explicitly the content of the speech and it is delivery will have the information about the fact that the audience is convinced enough to vote for or against him/her. Commonly if politician is polite but passionate enough to serve the people, talks about development, remain optimist in his speech, talks about facts and figures related to government policies to explain his point to the audience is expected to win and vice-versa. But we want to examine what does data say?
The response variable will indicate the win/loss information. The independent variables will be the characteristics of the speech which may affect the win/loss which are commonly the following.

1. Proportion of words in the speech showing *Optimism*
2. Proportion of words in the speech showing *Pessimism*
3. Proportion of words in the speech showing the use of *Past*
4. Proportion of words in the speech showing the use of *Present*
5. Proportion of words in the speech showing the use of *Future*
6. Number of time he/she mentions his/her own party
7. Number of time he/she mentions his/her opposite parties

8. Some measure indicating the content of speech showing *Openness*
9. Some measure indicating the content of speech showing *Conscientiousness*
10. Some measure indicating the content of speech showing *Extraversion*
11. Some measure indicating the content of speech showing *Agreeableness*
12. Some measure indicating the content of speech showing *Neuroticism*
13. Some measure indicating the content of speech showing *emotionality*

Case 4: Consider the US Presidential Data.xlsx and perform the following objectives:

What we are interested here in knowing that depending upon what and how a politician give speeches, his/her chances of winning the elections are affected

Step 1: Build a logistic regression model for classifying win/loss using the given independent variables

Logistic regression model for classifying win/loss using the given independent variables is given by:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Optimism	-3.568	2.090	2.913	1	.088	.028
	Pessimism	-28.452	2.951	92.951	1	.000	.000
	Past_Used	2.080	.763	7.438	1	.006	8.001
	Future_Used	4.138	.725	32.557	1	.000	62.665
	OwnPartyCount	.008	.006	1.757	1	.185	1.008
	OppPartyCount	.016	.013	1.466	1	.226	1.016
	NumericCount	311.997	53.759	33.682	1	.000	3.153E135
	Extra	-.377	.082	20.965	1	.000	.686
	Emoti	.212	.130	2.667	1	.102	1.236
	Agree	-.583	.188	9.645	1	.002	.558
	Consc	-.482	.118	16.771	1	.000	.618
	Openn	.828	.107	60.045	1	.000	2.288
	Constant	.936	.768	1.483	1	.223	2.549

a. Variable(s) entered on step 1: Optimism, Pessimism, Past_Used, Future_Used, OwnPartyCount, OppPartyCount, NumericCount, Extra, Emoti, Agree, Consc, Openn.

The fitted model is:

$$\pi(x) = \frac{e^{(0.936-3.568*X1-28.452*X2+2.080*X3+4.138*X4+0.008*X5+0.016*X6+311.997*X7-0.377*X8+0.212*X9-0.583*X10-0.492*X11+0.828*X12)}}{1+e^{(0.936-3.568*X1-28.452*X2+2.080*X3+4.138*X4+0.008*X5+0.016*X6+311.997*X7-0.377*X8+0.212*X9-0.583*X10-0.492*X11+0.828*X12)}}$$

We conclude that p-values of all the 12 variables i.e.$x_1,\ x_2, x_3,\ x_4,\ x_5,\ x_6,\ x_7,\ x_8,\ x_9,\ x_{10},\ x_{11},\ x_{12}$ are greater than 0.05 therefore all the independent variable are not significant at 5% level of significance.

Step 2: Test for the Significance individual independent variables.

Hosmer-Lemeshow Test of Goodness of Fit

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	43.903	8	.000

Contingency Table for Hosmer and Lemeshow Test

		Win_Loss = loss		Win_Loss = win		
		Observed	Expected	Observed	Expected	Total
Step 1	1	127	122.010	25	29.990	152
	2	113	102.257	39	49.743	152
	3	91	87.974	61	64.026	152
	4	57	74.678	95	77.322	152
	5	62	62.229	90	89.771	152
	6	55	49.828	97	102.172	152
	7	25	39.285	127	112.715	152
	8	28	29.436	124	122.564	152
	9	16	19.326	136	132.674	152
	10	21	7.978	135	148.022	156

H_o : Model is a good fit

H_1 : Model is not a good fit

INFERENCE :

Since the significant value of the test is 0.000 (<0.05) so null hypothesis gets rejected i.e. the model is not a good fit

Confusion or Classification Table

Classification Table[a]

			Predicted		
			Win_Loss		Percentage
	Observed		loss	win	Correct
Step 1	Win_Loss	loss	354	241	59.5
		win	157	772	83.1
	Overall Percentage				73.9

Classification Table[a]

Observed			Predicted		
			Win_Loss		Percentage Correct
			loss	win	
Step 1	Win_Loss	loss	354	241	59.5
		win	157	772	83.1
	Overall Percentage				73.9

a. The cut value is .500

INFERENCE :

Following **simple measures of goodness** can be defined.
1. Percentage of correct classification = 73.9% ≈ 74%
2. Sensitivity = 83.10%
3. Specificity = 59.4%

Skull Type Prediction using Discriminant Analysis

We are interested in predicting the type of skull of humans as one of two possible types I and II based on some five physical measures available related to the skulls and perform the following objectives.

1. Test for normality of all five physical measures.

We test this by Shapiro Wilk and Kolmogrov Smirnov test:
Kolmogorov-Smirnov & Shapiro-Wilk tests: These methods test whether one distribution (e.g. your dataset) is significantly different from another (e.g. a normal distribution) and produce a numerical answer, yes or no. Use the Shapiro-Wilk test if the sample size is between 3 and 2000 and the Kolmogorov-Smirnov test if the sample size is greater than 2000.

The Hypothesis which will be tested by the following two tests is

Ho: The distribution of the variable under consideration follows Normal Distribution. against
H1: The distribution of the variable under consideration does not follow Normal Distribution.

Kolmogorov-Smirnov Test for Normality: It is a non-parametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution. The

Kolmogorov–Smirnov statistic for a given cumulative distribution function F(x) is
Dn=sup$x|\ Fn\ (x)-\ F(x)|$
where $Fn(x)$ is the empirical distribution function computed using our sample.

Shapiro Wilk's W test: It is a test of normality in frequentist statistics. The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

One-Sample Kolmogorov-Smirnov Test

		X1	X2	X3	X4	X5
N		19	19	19	19	19
Normal Parameters[a]	Mean	53.4737	38.5263	47.8947	47.4211	55.0526
	Std. Deviation	3.44067E1	2.07588E1	3.02690E1	2.36909E1	3.29873E1
Most Extreme Differences	Absolute	.174	.112	.131	.127	.191
	Positive	.121	.090	.098	.102	.127
	Negative	-.174	-.112	-.131	-.127	-.191
Kolmogorov-Smirnov Z		.758	.487	.569	.553	.833
Asymp. Sig. (2-tailed)		.613	.972	.903	.920	.491

a. Test distribution is Normal.

The p values for all the statistics are greater than 0.05, which shows that all the 5 predictors can be taken to have normal distribution.

2. Test for equality of the covariance matrices of different groups using Box's M Test.

This test lets us test for the following hypothesis
Ho: Covariance matrices of the groups do not differ significantly. against
H1: Covariance matrices of the groups differ significantly.

Test Results

Box's M		21.637
F	Approx.	.947
	df1	15
	df2	908.297
	Sig.	.511

Tests null hypothesis of equal population covariance matrices.

Since, p-value is greater than 0.05. Hence we do not reject Ho at 5% level of significance and have an evidence that Covariance matrices of the groups may not differ significantly. Since this is a very powerful test, so when the sample is large, even small differences are considered significant. Thus, in order to be lenient, we check for values of Log Determinants. If these values for the groups are fairly close, we accept Ho and conclude that the covariance matrices are not significantly different.

Log Determinants

Skull_type	Rank	Log Determinant
0	5	32.677
1	5	29.947
Pooled within-groups	5	32.825

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Clearly, the values of log determinants of skull type 1 and skull type 2 do have some difference, hence we will reject Ho and conclude that covariance matrices of the groups differ significantly.

3. Test if the group means are statistically significantly different between the 2 groups for all the physical measures and point out which measures contribute to the discriminant function significantly.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
X1	.986	.242	1	17	.629
X2	.859	2.794	1	17	.113
X3	1.000	.005	1	17	.943
X4	.989	.184	1	17	.674
X5	.888	2.134	1	17	.162

We test the Hypothesis
Ho: All the group means are statistically not significantly different between the 2 groups for a particular measure.
*H*1: All the group means are statistically significantly different between the 2 groups for a particular measure.

As we know that value of Wilk's Lambda varies between 0 and 1. The smaller its value, the more the corresponding variable contributes to the discriminant function. Keeping this in mind, it can be seen from the table above that $X3$ has negligible contribution to discriminant function, whereas, $X2$ has the maximum contribution to the discriminant function.

4.To test the overall significance of discriminant function.

Let Hypothesis be defines as
Ho : The overall discriminant function is not significant.
H1: The overall discriminant function is significant.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.776	3.678	5	.597

Since p-value is greater than 0.05, we do not reject Ho at 5% level of significance, concluding that the discriminant function is not significant.Also the wilk's lambda is near to 1 so the discriminant analysis does not give a good fit.

5 .Obtain the standardized and unstandardised canonical discriminant

Standardized Canonical Discriminant Function Coefficients

	Function
	1
X1	.239
X2	.840
X3	.154
X4	.176
X5	-.592

Canonical Discriminant Function Coefficients

	Function
	1
X1	.007
X2	.042
X3	.005
X4	.007
X5	-.019
(Constant)	-1.559

Unstandardized coefficients

It Standardized canionical discriminant function coefficients are the variables rescaled to unit standard deviation. If a coefficient lies in the neighborhood of 1 or -1, then it is a good explanator & if it lies in the neighborhood of 0 or 0.5, then it gives a moderate explanation. SPSS generates Unstandardized Canonical Discriminant Function Coefficients as well which gives the same information but as they are not standardized the same rule of interpretation is not applicable.
So,from the first table of standardised canonical discriminant function coefficients the variables,viz., X1, X3, X4 and X5 are moderate explanators while X2 is a good explanatory.

6. Obtain the values of discriminant function at the group centroids and classification function coefficients.

Functions at Group Centroids

Skull_type	Function 1
0	.433
1	-.596

Unstandardized canonical discriminant functions evaluated at group means

From the above table the value of group centroids are:
m1 = .433 and m2 = -.596

7. Obtain the entries of the vector l, calculate the discriminant functions for the given skulls and hence obtain the predicted skull type for the given skulls.

Classification Function Coefficients

	Skull_type 0	Skull_type 1
X1	.100	.093
X2	.233	.189
X3	.055	.050
X4	.139	.132
X5	.062	.081
(Constant)	-14.614	-13.093

Fisher's linear discriminant functions

The value of l' is given in the table below:
l' = (0.007 0.044 0.005 0.007 -0.019)

The value of Discriminant function is: $l'X = 0.007*X1 + 0.044*X2 + 0.005*X3 + 0.007*X4 - 0.019*X5$

The predicted groups are given below in the table:

Skull Type	X1	X2	X3	X4	X5	Predicted	Discriminant
1	13	29	82	24	60	1	-0.77243772
0	79	1	1	56	63	1	-1.73441478
0	5	77	47	45	26	0	1.818153809
1	100	16	80	60	98	1	-1.18440226
0	55	65	3	20	2	0	1.694581665
0	91	55	20	59	68	0	0.660515664
1	47	31	31	52	19	0	0.253974788
1	17	43	61	45	79	1	-0.45430628
1	30	54	11	83	60	0	0.481433521
0	45	17	63	79	10	0	0.167127168
0	1	40	97	51	94	1	-0.74671213
1	69	44	40	47	84	1	-0.24014393
1	95	19	33	2	53	1	-0.91125726
0	83	47	75	68	9	0	1.695687271
0	78	57	86	19	88	0	0.322093423
1	94	1	50	44	88	1	-1.94051887
0	7	54	15	23	67	1	-0.22034139
0	77	45	34	32	75	1	-0.11528756
0	30	37	81	92	3	0	1.226254876

8. Construct the classification tables for fitting strength and predictive strength of the model. Also calculate sensitivity and specificity for both fitting strength and predictive strengths.

Classification Results[b,c]

		Skull_type	Predicted Group Membership		Total
			0	1	
Original	Count	0	7	4	11
		1	2	6	8
	%	0	63.6	36.4	100.0
		1	25.0	75.0	100.0
Cross-validated[a]	Count	0	5	6	11
		1	5	3	8
	%	0	45.5	54.5	100.0
		1	62.5	37.5	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 68.4% of original grouped cases correctly classified.

c. 42.1% of cross-validated grouped cases correctly classified.

For fitting strength: (based on the original data)
Sensitivity = 66+2 = 0.75 = 75%

Specificity = 77+4 = 0.6364 = 63.64%

For predicted strength:(Based on cross validated data)

Sensitivity = 33+5 = 0.375 = 37.5%

Specificity = 55+6 = 0.4546 = 45.46%

9. For a set of five physical measures given for a new skull in the dataset Skull Type Prediction – Validation Data.xlsx predict the skull type using the Logistic Classifier you have built.

X1	X2	X3	X4	X5	Predicted grps	Discriminant score
171	134	130	69	130	0	4.0228202344

The predicted group is 0.

Comparative Study of Binary Logistic Regression and Binary Discriminant Analysis:

The Skull Type Prediction Problem has already been solved using Logistic Regression and now you will be solving the same problem using a different technique Discriminant Analysis. Compare the two methodologies for Skull Type Prediction Problem on following grounds:

1. Classification of individual skull.

For binary logistic fitting the skull type data.

Skull Type	Logistic fitting	Discriminant	
1	1	1	TRUE
0	1	1	TRUE
0	0	0	TRUE
1	1	1	TRUE
0	0	0	TRUE
0	0	0	TRUE
1	0	0	TRUE
1	1	1	TRUE
1	0	0	TRUE
0	0	0	TRUE
0	1	1	TRUE
1	0	1	FALSE

1	1	1	TRUE
0	0	0	TRUE
0	0	0	TRUE
1	1	1	TRUE
0	0	1	FALSE
0	0	1	FALSE
0	0	0	TRUE

We will compare the predicted skull types in the case of binary logistic and discriminant analysis with the original skull types.

We can see that 3 cases in which the predicted groups are different by obtaining from both the methods. On comparing from the original skull type data in one case the model fit by discriminant analysis proves to be better as it gives the correct predicted value whereas in the other two cases model fitted by binary logistic proves to be better.

2. Confusion Matrix, Percentage of Correct Classification.

The classification table for binary logistic fitting:

Classification Table[a]

Observed		Predicted		
		Skull		
		0	1	Percentage Correct
Skull	**0**	9	2	81.8
	1	3	5	62.5
Overall Percentage				73.7

a. The cut value is .500

The classification table for discriminant analysis:

Classification Results[b,c]

		Skull_type	Predicted Group Membership		Total
			0	1	
Original	Count	0	7	4	11
		1	2	6	8
	%	0	63.6	36.4	100.0
		1	25.0	75.0	100.0
Cross-validated[a]	Count	0	5	6	11
		1	5	3	8
	%	0	45.5	54.5	100.0
		1	62.5	37.5	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 68.4% of original grouped cases correctly classified.

c. 42.1% of cross-validated grouped cases correctly classified.

The confusion matrices are given above and their correction percentage of classification are given as: (For the discriminant analysis classification table we will use the original data)

For discriminant analysis: 68.4%
For binary logistic: 73.7%

We will compare the diagonal entries i.e. the correctly classified ones (or the non diagonal elements i.e. the misclassified ones)
Clearly in one case 9 > 7, hence the binary logistic model is a good fit for skull type I, whereas 5 < 6 hence the discriminant analysis method fits better model for skull type II.

3. Sensitivity and Specificity

In case of logistic regression fitting:
Sensitivity= 5/(5+3) = 0.625 = 62.5%
Specificity = 10/(10+2) = 0.8334 = 83.34%

For discriminant analysis: For fitting strength:
Sensitivity = 6/(6+2) = 0.75 = 75%
Specificity = 7/(7+4) = 0.6364 = 63.64%

Since, more the sensitivity the better the model. So, the binary logistic model fit is better in terms sensitivity as the model fit by binary logistic method is more sensitive
The specificity of the model fitted by discriminant analysis is more, so it is a better model on the basis of specificity. The model obtained by discriminant analysis is more specific.

4. Performance on the Validation Data.

In logistic regression case

X1	X2	X3	X4	X5	Predicted prob	Grp Memb	Π_i	Expected
171	134	130	69	130	0.0067679492	0	0.0067679	0

In discriminant analysis case:

X1	X2	X3	X4	X5	Predicted grps	Discriminant score
171	134	130	69	130	0	4.0228202344

Both the method predict to be of Skull type 0.

Multiclass Classification

A classification problem is said to be multiclass classification problem if the response has more than two possible classes. We will be considering only Multiclass Logistic Regression as a multiclass classification technique though the other two techniques viz. Naïve Bayes' Classifier and Discriminant Analysis as well have their multivariate extensions.

Logistic Regression can be used to solve a multiclass classification problem in following two ways:

a) By means of decomposing the multiclass classification problem into several binary classification problems.
b) By means of using multinomial probability distribution.

Decomposing the multiclass classification problem into several binary classification problems:

Suppose the response variable has three class viz. 1, 2 and 3 then the response is defined as follows:

$$y_i = \begin{cases} 1 & \text{with prob } \pi_{1i} \\ 2 & \text{with prob } \pi_{2i} \\ 3 & \text{with prob } \pi_{3i} \end{cases}$$

We can define three binary variables using yi as follows:

$$y_{1i} = \begin{cases} 1 & \text{if } y_i = 1 \\ 0 & \text{otherwise} \end{cases}, \quad y_{2i} = \begin{cases} 1 & \text{if } y_i = 2 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad y_{3i} = \begin{cases} 1 & \text{if } y_i = 3 \\ 0 & \text{otherwise} \end{cases}$$

Clearly we have three binary classification problems which model $P[y_{1i} = 1] = \pi_{1i}$, $P[y_{2i} = 1] = \pi_{2i}$ and $P[y_{3i} = 1] = \pi_{3i}$ respectively.

Using Binary Logistic Regression we can obtain $\widehat{\pi_{1i}}, \widehat{\pi_{2i}}$ and $\widehat{\pi_{3i}}$ and can classify y_{1i}, y_{2i} and y_{3i} but our goal is to classify y_i the 3-class variable. We define the rule for multiclass classification as follows:

$$\hat{y}_i = \underset{k}{\operatorname{argmax}}\, \widehat{\pi_{ki}}$$

where $k = 1,2,3$ (# of classes) and $i = 1, \dots, n$ (# of observations).

Flower Species

Case 1: Considering data on Sepal Length, Sepal Width,Petal Length and Petal Width and Species Type for 150 different flowers and performing the following objectives.

1. Decompose the multiclass (3-class) classification problem into three Binary Classification Problems and perform the following for each problem:

a. Test for the Significance individual independent variables.

b. Test for the overall Logistic Regression using Hosmer and Lemeshow Test (It's a Chi-Square Test).

c. Construct the Confusion (Classification) Table and report the percentage of correct classification for the given emails. Also calculate specificity and sensitivity of the model.

d. For each flower obtain the predicted probability and hence the predicted class using the

Logistic Classifier you have built.

(i) Taking setosa as 1 ,versicolor and virginica as 0:

We define :
Ho: βi= 0 against H1: βi≠ 0 for atleast one of the i

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	sepal.length	8.666	1.484E4	.000	1	1.000	5.800E3
	sepal.width	6.637	6.923E3	.000	1	.999	762.489
	petal.length	-15.119	1.237E4	.000	1	.999	.000
	petal.width	-16.272	1.792E4	.000	1	.999	.000
	Constant	-13.616	5.186E4	.000	1	1.000	.000

a. Variable(s) entered on step 1: sepal.length, sepal.width, petal.length, petal.width.

From the table above, we cannot reject Ho in all of the above case.
We define :
Ho: $\beta = 0$ against H1: $\beta \neq 0$

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	8	1.000

Clearly we do not reject Ho.

Classification Table[a]

Observed		Predicted		
		Type		Percentage Correct
		0	1	
Step 1	Type 0	100	0	100.0
	1	0	50	100.0
	Overall Percentage			100.0

a. The cut value is .500

Sensitivity = 1.00 and specificity = 1.00

The predicted probabilities and predicted groups are:

Flower No.	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	Predicted probabilities	Predicted group
1	5.1	3.5	1.4	0.2	setosa	1	1
2	4.9	3	1.4	0.2	setosa	1	1
3	4.7	3.2	1.3	0.2	setosa	1	1

4	4.6	3.1	1.5	0.2	setosa	0.999999999	1
5	5	3.6	1.4	0.2	setosa	1	1
6	5.4	3.9	1.7	0.4	setosa	1	1
7	4.6	3.4	1.4	0.3	setosa	1	1
8	5	3.4	1.5	0.2	setosa	1	1
9	4.4	2.9	1.4	0.2	setosa	0.999999996	1
10	4.9	3.1	1.5	0.1	setosa	1	1
11	5.4	3.7	1.5	0.2	setosa	1	1
12	4.8	3.4	1.6	0.2	setosa	1	1
13	4.8	3	1.4	0.1	setosa	1	1
14	4.3	3	1.1	0.1	setosa	1	1
15	5.8	4	1.2	0.2	setosa	1	1
16	5.7	4.4	1.5	0.4	setosa	1	1
17	5.4	3.9	1.3	0.4	setosa	1	1
18	5.1	3.5	1.4	0.3	setosa	1	1
19	5.7	3.8	1.7	0.3	setosa	1	1
20	5.1	3.8	1.5	0.3	setosa	1	1
21	5.4	3.4	1.7	0.2	setosa	1	1
22	5.1	3.7	1.5	0.4	setosa	1	1
23	4.6	3.6	1	0.2	setosa	1	1
24	5.1	3.3	1.7	0.5	setosa	0.999999992	1
25	4.8	3.4	1.9	0.2	setosa	0.999999991	1
26	5	3	1.6	0.2	setosa	1	1
27	5	3.4	1.6	0.4	setosa	1	1
28	5.2	3.5	1.5	0.2	setosa	1	1
29	5.2	3.4	1.4	0.2	setosa	1	1
30	4.7	3.2	1.6	0.2	setosa	0.999999999	1
31	4.8	3.1	1.6	0.2	setosa	0.999999999	1
32	5.4	3.4	1.5	0.4	setosa	1	1
33	5.2	4.1	1.5	0.1	setosa	1	1
34	5.5	4.2	1.4	0.2	setosa	1	1
35	4.9	3.1	1.5	0.2	setosa	1	1
36	5	3.2	1.2	0.2	setosa	1	1
37	5.5	3.5	1.3	0.2	setosa	1	1
38	4.9	3.6	1.4	0.1	setosa	1	1
39	4.4	3	1.3	0.2	setosa	1	1
40	5.1	3.4	1.5	0.2	setosa	1	1
41	5	3.5	1.3	0.3	setosa	1	1
42	4.5	2.3	1.3	0.3	setosa	0.999999899	1
43	4.4	3.2	1.3	0.2	setosa	1	1
44	5	3.5	1.6	0.6	setosa	0.999999994	1
45	5.1	3.8	1.9	0.4	setosa	0.999999999	1
46	4.8	3	1.4	0.3	setosa	1	1
47	5.1	3.8	1.6	0.2	setosa	1	1
48	4.6	3.2	1.4	0.2	setosa	1	1
49	5.3	3.7	1.5	0.2	setosa	1	1

50	5	3.3	1.4	0.2	setosa	1	1
51	7	3.2	4.7	1.4	versicolor	7.94E-12	0
52	6.4	3.2	4.5	1.5	versicolor	1.77E-13	0
53	6.9	3.1	4.9	1.5	versicolor	1.64E-14	0
54	5.5	2.3	4	1.3	versicolor	9.20E-15	0
55	6.5	2.8	4.6	1.5	versicolor	6.53E-15	0
56	5.7	2.8	4.5	1.3	versicolor	7.49E-16	0
57	6.3	3.3	4.7	1.6	versicolor	1.38E-15	0
58	4.9	2.4	3.3	1	versicolor	5.13E-10	0
59	6.6	2.9	4.6	1.3	versicolor	7.82E-13	0
60	5.2	2.7	3.9	1.4	versicolor	8.66E-15	0
61	5	2	3.5	1	versicolor	4.17E-12	0
62	5.9	3	4.2	1.5	versicolor	5.75E-14	0
63	6	2.2	4	1	versicolor	4.76E-11	0
64	6.1	2.9	4.7	1.4	versicolor	4.45E-16	0
65	5.6	2.9	3.6	1.3	versicolor	4.96E-10	0
66	6.7	3.1	4.4	1.4	versicolor	2.83E-11	0
67	5.6	3	4.5	1.5	versicolor	4.58E-17	0
68	5.8	2.7	4.1	1	versicolor	5.12E-11	0
69	6.2	2.2	4.5	1.5	versicolor	4.11E-17	0
70	5.6	2.5	3.9	1.1	versicolor	9.69E-12	0
71	5.9	3.2	4.8	1.8	versicolor	1.89E-19	0
72	6.1	2.8	4	1.3	versicolor	4.60E-11	0
73	6.3	2.5	4.9	1.5	versicolor	1.69E-18	0
74	6.1	2.8	4.7	1.2	versicolor	5.93E-15	0
75	6.4	2.9	4.3	1.3	versicolor	1.29E-11	0
76	6.6	3	4.4	1.4	versicolor	6.14E-12	0
77	6.8	2.8	4.8	1.4	versicolor	2.18E-14	0
78	6.7	3	5	1.7	versicolor	1.27E-17	0
79	6	2.9	4.5	1.5	versicolor	7.56E-16	0
80	5.7	2.6	3.5	1	versicolor	9.64E-08	0
81	5.5	2.4	3.8	1.1	versicolor	9.52E-12	0
82	5.5	2.4	3.7	1	versicolor	2.20E-10	0
83	5.8	2.7	3.9	1.2	versicolor	4.06E-11	0
84	6	2.7	5.1	1.6	versicolor	4.52E-21	0
85	5.4	3	4.5	1.5	versicolor	8.10E-18	0
86	6	3.4	4.5	1.6	versicolor	4.10E-15	0
87	6.7	3.1	4.7	1.5	versicolor	5.97E-14	0
88	6.3	2.3	4.4	1.3	versicolor	2.23E-14	0
89	5.6	3	4.1	1.3	versicolor	5.02E-13	0
90	5.5	2.5	4	1.3	versicolor	3.47E-14	0
91	5.5	2.6	4.4	1.2	versicolor	8.10E-16	0
92	6.1	3	4.6	1.4	versicolor	3.92E-15	0
93	5.8	2.6	4	1.2	versicolor	4.61E-12	0
94	5	2.3	3.3	1	versicolor	6.28E-10	0
95	5.6	2.7	4.2	1.3	versicolor	1.51E-14	0

96	5.7	3	4.2	1.2	versicolor	1.34E-12	0
97	5.7	2.9	4.2	1.3	versicolor	1.36E-13	0
98	6.2	2.9	4.3	1.3	versicolor	2.28E-12	0
99	5.1	2.5	3	1.1	versicolor	1.03E-07	0
100	5.7	2.8	4.1	1.3	versicolor	3.17E-13	0
101	6.3	3.3	6	2.5	virginica	1.75E-30	0
102	5.8	2.7	5.1	1.9	virginica	6.06E-24	0
103	7.1	3	5.9	2.1	virginica	7.47E-25	0
104	6.3	2.9	5.6	1.8	virginica	4.62E-24	0
105	6.5	3	5.8	2.2	virginica	3.68E-27	0
106	7.6	3	6.6	2.1	virginica	1.44E-27	0
107	4.9	2.5	4.5	1.7	virginica	1.49E-22	0
108	7.3	2.9	6.3	1.8	virginica	6.79E-25	0
109	6.7	2.5	5.8	1.8	virginica	5.06E-25	0
110	7.2	3.6	6.1	2.5	virginica	6.91E-27	0
111	6.5	3.2	5.1	2	virginica	1.42E-20	0
112	6.4	2.7	5.3	1.9	virginica	5.34E-23	0
113	6.8	3	5.5	2.1	virginica	2.35E-23	0
114	5.7	2.5	5	2	virginica	6.02E-25	0
115	5.8	2.8	5.1	2.4	virginica	3.45E-27	0
116	6.4	3.2	5.3	2.3	virginica	2.20E-24	0
117	6.5	3	5.5	1.8	virginica	2.30E-22	0
118	7.7	3.8	6.7	2.2	virginica	3.01E-26	0
119	7.7	2.6	6.9	2.3	virginica	9.98E-32	0
120	6	2.2	5	1.5	virginica	3.78E-21	0
121	6.9	3.2	5.7	2.3	virginica	3.96E-25	0
122	5.6	2.8	4.9	2	virginica	8.41E-24	0
123	7.7	2.8	6.7	2	virginica	1.02E-27	0
124	6.3	2.7	4.9	1.8	virginica	4.83E-20	0
125	6.7	3.3	5.7	2.1	virginica	3.52E-24	0
126	7.2	3.2	6	1.8	virginica	1.95E-22	0
127	6.2	2.8	4.8	1.8	virginica	1.79E-19	0
128	6.1	3	4.9	1.8	virginica	6.26E-20	0
129	6.4	2.8	5.6	2.1	virginica	4.29E-26	0
130	7.2	3	5.8	1.6	virginica	2.75E-20	0
131	7.4	2.8	6.1	1.9	virginica	3.36E-24	0
132	7.9	3.8	6.4	2	virginica	4.11E-22	0
133	6.4	2.8	5.6	2.2	virginica	8.43E-27	0
134	6.3	2.8	5.1	1.5	virginica	6.02E-19	0
135	6.1	2.6	5.6	1.4	virginica	7.48E-23	0
136	7.7	3	6.1	2.3	virginica	2.54E-25	0
137	6.3	3.4	5.6	2.4	virginica	7.34E-27	0
138	6.4	3.1	5.5	1.8	virginica	1.88E-22	0
139	6	3	4.8	1.8	virginica	1.19E-19	0
140	6.9	3.1	5.4	2.1	virginica	4.92E-22	0
141	6.7	3.1	5.6	2.4	virginica	3.21E-26	0

142	6.9	3.1	5.1	2.3	virginica	1.77E-21	0
143	5.8	2.7	5.1	1.9	virginica	6.06E-24	0
144	6.8	3.2	5.9	2.3	virginica	8.08E-27	0
145	6.7	3.3	5.7	2.5	virginica	5.24E-27	0
146	6.7	3	5.2	2.3	virginica	3.56E-23	0
147	6.3	2.5	5	1.9	virginica	5.55E-22	0
148	6.5	3	5.2	2	virginica	8.29E-22	0
149	6.2	3.4	5.4	2.3	virginica	3.23E-25	0
150	5.9	3	5.1	1.8	virginica	5.38E-22	0

There is no misprediction in this case.

(ii) **Taking versicolor as 1 ,setosa and virginica as 0:**

We define : Ho: $\beta i = 0$ against H1: $\beta i \neq 0$ for atleast one of the i

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	sepal.length	-.245	.650	.143	1	.706	.782
	sepal.width	-2.797	.784	12.739	1	.000	.061
	petal.length	1.314	.684	3.691	1	.055	3.720
	petal.width	-2.778	1.173	5.609	1	.018	.062
	Constant	7.378	2.499	8.716	1	.003	1.601E3

a. Variable(s) entered on step 1: sepal.length, sepal.width, petal.length, petal.width.

We see that we do not reject Ho in case of coefficients of sepal_length and petal_length while we reject *Ho* for other variables .

We define : Ho: $\beta = 0$ against H1: $\beta \neq 0$.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8.524	8	.384

Clearly, since p-value is greater than 0.05, we do not reject Ho.

Classification Table[a]

Observed			Predicted		
			Type		Percentage Correct
			0	1	
Step 1	Type	0	86	14	86.0
		1	25	25	50.0
	Overall Percentage				74.0

a. The cut value is .500

Sensitivity= 0.50 and Specificity = 0.86

The predicted probabilities and predicted groups are:

Flower No.	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	Predicted probabilities	Predicted group
1	5.1	3.5	1.4	0.2	0	0.084913	0
2	4.9	3	1.4	0.2	0	0.282918	0
3	4.7	3.2	1.3	0.2	0	0.171983	0
4	4.6	3.1	1.5	0.2	0	0.268015	0
5	5	3.6	1.4	0.2	0	0.067075	0
6	5.4	3.9	1.7	0.4	0	0.023403	0
7	4.6	3.4	1.4	0.3	0	0.095101	0
8	5	3.4	1.5	0.2	0	0.125447	0
9	4.4	2.9	1.4	0.2	0	0.371054	0
10	4.9	3.1	1.5	0.1	0	0.30992	0
11	5.4	3.7	1.5	0.2	0	0.053204	0
12	4.8	3.4	1.6	0.2	0	0.146616	0
13	4.8	3	1.4	0.1	0	0.34804	0
14	4.3	3	1.1	0.1	0	0.28924	0
15	5.8	4	1.2	0.2	0	0.014627	0
16	5.7	4.4	1.5	0.4	0	0.004211	0
17	5.4	3.9	1.3	0.4	0	0.013972	0
18	5.1	3.5	1.4	0.3	0	0.065668	0
19	5.7	3.8	1.7	0.3	0	0.037423	0
20	5.1	3.8	1.5	0.3	0	0.033478	0
21	5.4	3.4	1.7	0.2	0	0.144643	0
22	5.1	3.7	1.5	0.4	0	0.033537	0
23	4.6	3.6	1	0.2	0	0.044795	0
24	5.1	3.3	1.7	0.5	0	0.094706	0
25	4.8	3.4	1.9	0.2	0	0.203056	0
26	5	3	1.6	0.2	0	0.333624	0

27	5	3.4	1.6	0.4	0	0.085792	0
28	5.2	3.5	1.5	0.2	0	0.093591	0
29	5.2	3.4	1.4	0.2	0	0.106951	0
30	4.7	3.2	1.6	0.2	0	0.235493	0
31	4.8	3.1	1.6	0.2	0	0.284464	0
32	5.4	3.4	1.5	0.4	0	0.069419	0
33	5.2	4.1	1.5	0.1	0	0.024827	0
34	5.5	4.2	1.4	0.2	0	0.011738	0
35	4.9	3.1	1.5	0.2	0	0.253823	0
36	5	3.2	1.2	0.2	0	0.144722	0
37	5.5	3.5	1.3	0.2	0	0.068696	0
38	4.9	3.6	1.4	0.1	0	0.088657	0
39	4.4	3	1.3	0.2	0	0.281159	0
40	5.1	3.4	1.5	0.2	0	0.12278	0
41	5	3.5	1.3	0.3	0	0.05941	0
42	4.5	2.3	1.3	0.3	0	0.671838	1
43	4.4	3.2	1.3	0.2	0	0.182719	0
44	5	3.5	1.6	0.6	0	0.039111	0
45	5.1	3.8	1.9	0.4	0	0.042484	0
46	4.8	3	1.4	0.3	0	0.234454	0
47	5.1	3.8	1.6	0.2	0	0.049565	0
48	4.6	3.2	1.4	0.2	0	0.19533	0
49	5.3	3.7	1.5	0.2	0	0.054454	0
50	5	3.3	1.4	0.2	0	0.142639	0
51	7	3.2	4.7	1.4	1	0.268237	0
52	6.4	3.2	4.5	1.5	1	0.198303	0
53	6.9	3.1	4.9	1.5	1	0.328605	0
54	5.5	2.3	4	1.3	1	0.775502	1
55	6.5	2.8	4.6	1.5	1	0.457235	0
56	5.7	2.8	4.5	1.3	1	0.610428	1
57	6.3	3.3	4.7	1.6	1	0.158803	0
58	4.9	2.4	3.3	1	1	0.735197	1
59	6.6	2.9	4.6	1.3	1	0.519989	1
60	5.2	2.7	3.9	1.4	1	0.446561	0
61	5	2	3.5	1	1	0.915132	1
62	5.9	3	4.2	1.5	1	0.248051	0
63	6	2.2	4	1	1	0.902922	1
64	6.1	2.9	4.7	1.4	1	0.514048	1
65	5.6	2.9	3.6	1.3	1	0.27125	0
66	6.7	3.1	4.4	1.4	1	0.260299	0
67	5.6	3	4.5	1.5	1	0.344945	0
68	5.8	2.7	4.1	1	1	0.733466	1
69	6.2	2.2	4.5	1.5	1	0.809799	1
70	5.6	2.5	3.9	1.1	1	0.746513	1
71	5.9	3.2	4.8	1.8	1	0.152689	0
72	6.1	2.8	4	1.3	1	0.424127	0

73	6.3	2.5	4.9	1.5	1	0.752251	1
74	6.1	2.8	4.7	1.2	1	0.709206	1
75	6.4	2.9	4.3	1.3	1	0.434129	0
76	6.6	3	4.4	1.4	1	0.322955	0
77	6.8	2.8	4.8	1.4	1	0.573336	1
78	6.7	3	5	1.7	1	0.307874	0
79	6	2.9	4.5	1.5	1	0.387028	0
80	5.7	2.6	3.5	1	1	0.629083	1
81	5.5	2.4	3.8	1.1	1	0.777806	1
82	5.5	2.4	3.7	1	1	0.802088	1
83	5.8	2.7	3.9	1.2	1	0.548319	1
84	6	2.7	5.1	1.6	1	0.647905	1
85	5.4	3	4.5	1.5	1	0.356115	0
86	6	3.4	4.5	1.6	1	0.105652	0
87	6.7	3.1	4.7	1.5	1	0.283299	0
88	6.3	2.3	4.4	1.3	1	0.827615	1
89	5.6	3	4.1	1.3	1	0.351803	0
90	5.5	2.5	4	1.3	1	0.663812	1
91	5.5	2.6	4.4	1.2	1	0.769229	1
92	6.1	3	4.6	1.4	1	0.412215	0
93	5.8	2.6	4	1.2	1	0.646777	1
94	5	2.3	3.3	1	1	0.781815	1
95	5.6	2.7	4.2	1.3	1	0.588849	1
96	5.7	3	4.2	1.2	1	0.443625	0
97	5.7	2.9	4.2	1.3	1	0.444075	0
98	6.2	2.9	4.3	1.3	1	0.446221	0
99	5.1	2.5	3	1.1	1	0.505125	1
100	5.7	2.8	4.1	1.3	1	0.480923	0
101	6.3	3.3	6	2.5	0	0.078715	0
102	5.8	2.7	5.1	1.9	0	0.456463	0
103	7.1	3	5.9	2.1	0	0.302109	0
104	6.3	2.9	5.6	1.8	0	0.519509	1
105	6.5	3	5.8	2.2	0	0.249878	0
106	7.6	3	6.6	2.1	0	0.4899	0
107	4.9	2.5	4.5	1.7	0	0.592191	1
108	7.3	2.9	6.3	1.8	0	0.679673	1
109	6.7	2.5	5.8	1.8	0	0.795969	1
110	7.2	3.6	6.1	2.5	0	0.032661	0
111	6.5	3.2	5.1	2	0	0.116865	0
112	6.4	2.7	5.3	1.9	0	0.485236	0
113	6.8	3	5.5	2.1	0	0.216001	0
114	5.7	2.5	5	2	0	0.500015	1
115	5.8	2.8	5.1	2.4	0	0.136648	0
116	6.4	3.2	5.3	2.3	0	0.07118	0
117	6.5	3	5.5	1.8	0	0.40564	0
118	7.7	3.8	6.7	2.2	0	0.079536	0

119	7.7	2.6	6.9	2.3	0	0.709326	1
120	6	2.2	5	1.5	0	0.896098	1
121	6.9	3.2	5.7	2.3	0	0.102853	0
122	5.6	2.8	4.9	2	0	0.279743	0
123	7.7	2.8	6.7	2	0	0.711683	1
124	6.3	2.7	4.9	1.8	0	0.429924	0
125	6.7	3.3	5.7	2.1	0	0.136953	0
126	7.2	3.2	6	1.8	0	0.387874	0
127	6.2	2.8	4.8	1.8	0	0.3388	0
128	6.1	3	4.9	1.8	0	0.255011	0
129	6.4	2.8	5.6	2.1	0	0.377469	0
130	7.2	3	5.8	1.6	0	0.597725	1
131	7.4	2.8	6.1	1.9	0	0.614632	1
132	7.9	3.8	6.4	2	0	0.088172	0
133	6.4	2.8	5.6	2.2	0	0.314721	0
134	6.3	2.8	5.1	1.5	0	0.630512	1
135	6.1	2.6	5.6	1.4	0	0.888693	1
136	7.7	3	6.1	2.3	0	0.217988	0
137	6.3	3.4	5.6	2.4	0	0.048007	0
138	6.4	3.1	5.5	1.8	0	0.345894	0
139	6	3	4.8	1.8	0	0.235252	0
140	6.9	3.1	5.4	2.1	0	0.151269	0
141	6.7	3.1	5.6	2.4	0	0.095662	0
142	6.9	3.1	5.1	2.3	0	0.064498	0
143	5.8	2.7	5.1	1.9	0	0.456463	0
144	6.8	3.2	5.9	2.3	0	0.132544	0
145	6.7	3.3	5.7	2.5	0	0.049634	0
146	6.7	3	5.2	2.3	0	0.098469	0
147	6.3	2.5	5	1.9	0	0.532624	1
148	6.5	3	5.2	2	0	0.208865	0
149	6.2	3.4	5.4	2.3	0	0.049851	0
150	5.9	3	5.1	1.8	0	0.31859	0

In the data there are 39 mispredicted values.

(iii) **Taking virginica as 1, setosa and versicolor as 0:**

We define :

Ho: $\beta i = 0$ against H1: $\beta i \neq 0$ for atleast one of the i

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	sepal.length	-2.465	2.394	1.060	1	.303	.085
	sepal.width	-6.681	4.480	2.224	1	.136	.001
	petal.length	9.429	4.737	3.962	1	.047	1.245E4

	petal.width	18.286	9.743	3.523	1	.061	8.741E7
	Constant	-42.638	25.708	2.751	1	.097	.000

a. Variable(s) entered on step 1: sepal.length, sepal.width, petal.length, petal.width.

Clearly, we do not reject Ho for coefficients corresponding to sepal_length, sepal_width, petal_width and for constant whereas reject Ho for petal_length.

We define :

Ho: $\beta = 0$ against H1: $\beta \neq 0$

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.259	8	1.000

Clearly, since p- value is greater than 0.05, hence we do not reject Ho.

Classification Table[a]

Observed		Predicted		
		Type		Percentage Correct
		0	1	
Step 1	Type 0	99	1	99.0
	1	1	49	98.0
	Overall Percentage			98.7

a. The cut value is .500

Sensitivity =0.98 and Specificity = 0.99
The predicted probabilities and predicted groups are:

Flower No.	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	Predicted probabilities	Predicted group
1	5.1	3.5	1.4	0.2	0	1.54E-27	0
2	4.9	3	1.4	0.2	0	7.14E-26	0
3	4.7	3.2	1.3	0.2	0	1.20E-26	0
4	4.6	3.1	1.5	0.2	0	1.97E-25	0
5	5	3.6	1.4	0.2	0	1.01E-27	0
6	5.4	3.9	1.7	0.4	0	3.34E-26	0
7	4.6	3.4	1.4	0.3	0	6.43E-26	0
8	5	3.4	1.5	0.2	0	9.90E-27	0
9	4.4	2.9	1.4	0.2	0	4.78E-25	0
10	4.9	3.1	1.5	0.1	0	1.51E-26	0
11	5.4	3.7	1.5	0.2	0	4.97E-28	0

12	4.8	3.4	1.6	0.2	0	4.16E-26	0
13	4.8	3	1.4	0.1	0	1.47E-26	0
14	4.3	3	1.1	0.1	0	2.97E-27	0
15	5.8	4	1.2	0.2	0	1.48E-30	0
16	5.7	4.4	1.5	0.4	0	8.57E-29	0
17	5.4	3.9	1.3	0.4	0	7.69E-28	0
18	5.1	3.5	1.4	0.3	0	9.61E-27	0
19	5.7	3.8	1.7	0.3	0	5.00E-27	0
20	5.1	3.8	1.5	0.3	0	3.33E-27	0
21	5.4	3.4	1.7	0.2	0	2.43E-26	0
22	5.1	3.7	1.5	0.4	0	4.04E-26	0
23	4.6	3.6	1	0.2	0	6.25E-29	0
24	5.1	3.3	1.7	0.5	0	2.40E-23	0
25	4.8	3.4	1.9	0.2	0	7.04E-25	0
26	5	3	1.6	0.2	0	3.68E-25	0
27	5	3.4	1.6	0.4	0	9.85E-25	0
28	5.2	3.5	1.5	0.2	0	3.10E-27	0
29	5.2	3.4	1.4	0.2	0	2.35E-27	0
30	4.7	3.2	1.6	0.2	0	2.03E-25	0
31	4.8	3.1	1.6	0.2	0	3.09E-25	0
32	5.4	3.4	1.5	0.4	0	1.43E-25	0
33	5.2	4.1	1.5	0.1	0	9.04E-30	0
34	5.5	4.2	1.4	0.2	0	5.36E-30	0
35	4.9	3.1	1.5	0.2	0	9.40E-26	0
36	5	3.2	1.2	0.2	0	2.22E-27	0
37	5.5	3.5	1.3	0.2	0	2.24E-28	0
38	4.9	3.6	1.4	0.1	0	2.08E-28	0
39	4.4	3	1.3	0.2	0	9.54E-26	0
40	5.1	3.4	1.5	0.2	0	7.73E-27	0
41	5	3.5	1.3	0.3	0	4.79E-27	0
42	4.5	2.3	1.3	0.3	0	4.98E-23	0
43	4.4	3.2	1.3	0.2	0	2.51E-26	0
44	5	3.5	1.6	0.6	0	1.96E-23	0
45	5.1	3.8	1.9	0.4	0	9.00E-25	0
46	4.8	3	1.4	0.3	0	5.69E-25	0
47	5.1	3.8	1.6	0.2	0	1.37E-27	0
48	4.6	3.2	1.4	0.2	0	3.93E-26	0
49	5.3	3.7	1.5	0.2	0	6.37E-28	0
50	5	3.3	1.4	0.2	0	7.52E-27	0
51	7	3.2	4.7	1.4	0	1.17E-05	0
52	6.4	3.2	4.5	1.5	0	4.86E-05	0
53	6.9	3.1	4.9	1.5	0	0.001199	0
54	5.5	2.3	4	1.3	0	4.22E-05	0
55	6.5	2.8	4.6	1.5	0	0.001408	0
56	5.7	2.8	4.5	1.3	0	1.02E-04	0
57	6.3	3.3	4.7	1.6	0	0.001306	0

58	4.9	2.4	3.3	1	0	5.35E-10	0
59	6.6	2.9	4.6	1.3	0	1.46E-05	0
60	5.2	2.7	3.9	1.4	0	1.48E-05	0
61	5	2	3.5	1	0	3.99E-08	0
62	5.9	3	4.2	1.5	0	3.74E-05	0
63	6	2.2	4	1	0	9.95E-08	0
64	6.1	2.9	4.7	1.4	0	7.99E-04	0
65	5.6	2.9	3.6	1.3	0	1.38E-08	0
66	6.7	3.1	4.4	1.4	0	2.83E-06	0
67	5.6	3	4.5	1.5	0	0.001326	0
68	5.8	2.7	4.1	1	0	1.48E-08	0
69	6.2	2.2	4.5	1.5	0	0.059598	0
70	5.6	2.5	3.9	1.1	0	8.71E-08	0
71	5.9	3.2	4.8	1.8	0	0.404838	0
72	6.1	2.8	4	1.3	0	3.41E-07	0
73	6.3	2.5	4.9	1.5	0	0.224834	0
74	6.1	2.8	4.7	1.2	0	4.02E-05	0
75	6.4	2.9	4.3	1.3	0	1.41E-06	0
76	6.6	3	4.4	1.4	0	7.06E-06	0
77	6.8	2.8	4.8	1.4	0	7.12E-04	0
78	6.7	3	5	1.7	0	0.276062	0
79	6	2.9	4.5	1.5	0	9.65E-04	0
80	5.7	2.6	3.5	1	0	1.29E-10	0
81	5.5	2.4	3.8	1.1	0	8.47E-08	0
82	5.5	2.4	3.7	1	0	5.30E-09	0
83	5.8	2.7	3.9	1.2	0	8.71E-08	0
84	6	2.7	5.1	1.6	0	0.86763	1
85	5.4	3	4.5	1.5	0	0.002169	0
86	6	3.4	4.5	1.6	0	2.13E-04	0
87	6.7	3.1	4.7	1.5	0	2.98E-04	0
88	6.3	2.3	4.4	1.3	0	2.55E-04	0
89	5.6	3	4.1	1.3	0	7.88E-07	0
90	5.5	2.5	4	1.3	0	1.11E-05	0
91	5.5	2.6	4.4	1.2	0	3.97E-05	0
92	6.1	3	4.6	1.4	0	1.60E-04	0
93	5.8	2.6	4	1.2	0	4.36E-07	0
94	5	2.3	3.3	1	0	8.16E-10	0
95	5.6	2.7	4.2	1.3	0	1.50E-05	0
96	5.7	3	4.2	1.2	0	2.54E-07	0
97	5.7	2.9	4.2	1.3	0	3.09E-06	0
98	6.2	2.9	4.3	1.3	0	2.31E-06	0
99	5.1	2.5	3	1.1	0	6.16E-11	0
100	5.7	2.8	4.1	1.3	0	2.34E-06	0
101	6.3	3.3	6	2.5	1	1	1
102	5.8	2.7	5.1	1.9	1	0.999614	1
103	7.1	3	5.9	2.1	1	0.999999	1

104	6.3	2.9	5.6	1.8	1	0.999719	1
105	6.5	3	5.8	2.2	1	1	1
106	7.6	3	6.6	2.1	1	1	1
107	4.9	2.5	4.5	1.7	1	0.890812	1
108	7.3	2.9	6.3	1.8	1	0.999996	1
109	6.7	2.5	5.8	1.8	1	0.999992	1
110	7.2	3.6	6.1	2.5	1	1	1
111	6.5	3.2	5.1	2	1	0.990258	1
112	6.4	2.7	5.3	1.9	1	0.999743	1
113	6.8	3	5.5	2.1	1	0.99998	1
114	5.7	2.5	5	2	1	0.999967	1
115	5.8	2.8	5.1	2.4	1	1	1
116	6.4	3.2	5.3	2.3	1	0.999995	1
117	6.5	3	5.5	1.8	1	0.997699	1
118	7.7	3.8	6.7	2.2	1	1	1
119	7.7	2.6	6.9	2.3	1	1	1
120	6	2.2	5	1.5	1	0.920492	1
121	6.9	3.2	5.7	2.3	1	1	1
122	5.6	2.8	4.9	2	1	0.999513	1
123	7.7	2.8	6.7	2	1	1	1
124	6.3	2.7	4.9	1.8	1	0.948434	1
125	6.7	3.3	5.7	2.1	1	0.999982	1
126	7.2	3.2	6	1.8	1	0.999559	1
127	6.2	2.8	4.8	1.8	1	0.824544	1
128	6.1	3	4.9	1.8	1	0.802299	1
129	6.4	2.8	5.6	2.1	1	0.999999	1
130	7.2	3	5.8	1.6	1	0.971201	1
131	7.4	2.8	6.1	1.9	1	0.999997	1
132	7.9	3.8	6.4	2	1	0.999919	1
133	6.4	2.8	5.6	2.2	1	1	1
134	6.3	2.8	5.1	1.5	1	0.204874	0
135	6.1	2.6	5.6	1.4	1	0.966405	1
136	7.7	3	6.1	2.3	1	1	1
137	6.3	3.4	5.6	2.4	1	1	1
138	6.4	3.1	5.5	1.8	1	0.996497	1
139	6	3	4.8	1.8	1	0.669142	1
140	6.9	3.1	5.4	2.1	1	0.999872	1
141	6.7	3.1	5.6	2.4	1	1	1
142	6.9	3.1	5.1	2.3	1	0.999944	1
143	5.8	2.7	5.1	1.9	1	0.999614	1
144	6.8	3.2	5.9	2.3	1	1	1
145	6.7	3.3	5.7	2.5	1	1	1
146	6.7	3	5.2	2.3	1	0.999993	1
147	6.3	2.5	5	1.9	1	0.999107	1
148	6.5	3	5.2	2	1	0.998994	1
149	6.2	3.4	5.4	2.3	1	0.999996	1

150	5.9	3	5.1	1.8	1	0.977679	1

In the data of 150 there are 2 mispredictions.

2. Obtain the multiclass predicted flower species for the original problem using the three sub problems.

Flower No.	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	y	Predicted group
1	5.1	3.5	1.4	0.2	setosa	1	1
2	4.9	3	1.4	0.2	setosa	1	1
3	4.7	3.2	1.3	0.2	setosa	1	1
4	4.6	3.1	1.5	0.2	setosa	1	1
5	5	3.6	1.4	0.2	setosa	1	1
6	5.4	3.9	1.7	0.4	setosa	1	1
7	4.6	3.4	1.4	0.3	setosa	1	1
8	5	3.4	1.5	0.2	setosa	1	1
9	4.4	2.9	1.4	0.2	setosa	1	1
10	4.9	3.1	1.5	0.1	setosa	1	1
11	5.4	3.7	1.5	0.2	setosa	1	1
12	4.8	3.4	1.6	0.2	setosa	1	1
13	4.8	3	1.4	0.1	setosa	1	1
14	4.3	3	1.1	0.1	setosa	1	1
15	5.8	4	1.2	0.2	setosa	1	1
16	5.7	4.4	1.5	0.4	setosa	1	1
17	5.4	3.9	1.3	0.4	setosa	1	1
18	5.1	3.5	1.4	0.3	setosa	1	1
19	5.7	3.8	1.7	0.3	setosa	1	1
20	5.1	3.8	1.5	0.3	setosa	1	1
21	5.4	3.4	1.7	0.2	setosa	1	1
22	5.1	3.7	1.5	0.4	setosa	1	1
23	4.6	3.6	1	0.2	setosa	1	1
24	5.1	3.3	1.7	0.5	setosa	1	1
25	4.8	3.4	1.9	0.2	setosa	1	1
26	5	3	1.6	0.2	setosa	1	1
27	5	3.4	1.6	0.4	setosa	1	1
28	5.2	3.5	1.5	0.2	setosa	1	1
29	5.2	3.4	1.4	0.2	setosa	1	1
30	4.7	3.2	1.6	0.2	setosa	1	1
31	4.8	3.1	1.6	0.2	setosa	1	1
32	5.4	3.4	1.5	0.4	setosa	1	1
33	5.2	4.1	1.5	0.1	setosa	1	1
34	5.5	4.2	1.4	0.2	setosa	1	1
35	4.9	3.1	1.5	0.2	setosa	1	1
36	5	3.2	1.2	0.2	setosa	1	1
37	5.5	3.5	1.3	0.2	setosa	1	1
38	4.9	3.6	1.4	0.1	setosa	1	1

39	4.4	3	1.3	0.2	setosa	1	1
40	5.1	3.4	1.5	0.2	setosa	1	1
41	5	3.5	1.3	0.3	setosa	1	1
42	4.5	2.3	1.3	0.3	setosa	1	1
43	4.4	3.2	1.3	0.2	setosa	1	1
44	5	3.5	1.6	0.6	setosa	1	1
45	5.1	3.8	1.9	0.4	setosa	1	1
46	4.8	3	1.4	0.3	setosa	1	1
47	5.1	3.8	1.6	0.2	setosa	1	1
48	4.6	3.2	1.4	0.2	setosa	1	1
49	5.3	3.7	1.5	0.2	setosa	1	1
50	5	3.3	1.4	0.2	setosa	1	1
51	7	3.2	4.7	1.4	versicolor	2	2
52	6.4	3.2	4.5	1.5	versicolor	2	2
53	6.9	3.1	4.9	1.5	versicolor	2	2
54	5.5	2.3	4	1.3	versicolor	2	2
55	6.5	2.8	4.6	1.5	versicolor	2	2
56	5.7	2.8	4.5	1.3	versicolor	2	2
57	6.3	3.3	4.7	1.6	versicolor	2	2
58	4.9	2.4	3.3	1	versicolor	2	2
59	6.6	2.9	4.6	1.3	versicolor	2	2
60	5.2	2.7	3.9	1.4	versicolor	2	2
61	5	2	3.5	1	versicolor	2	2
62	5.9	3	4.2	1.5	versicolor	2	2
63	6	2.2	4	1	versicolor	2	2
64	6.1	2.9	4.7	1.4	versicolor	2	2
65	5.6	2.9	3.6	1.3	versicolor	2	2
66	6.7	3.1	4.4	1.4	versicolor	2	2
67	5.6	3	4.5	1.5	versicolor	2	2
68	5.8	2.7	4.1	1	versicolor	2	2
69	6.2	2.2	4.5	1.5	versicolor	2	2
70	5.6	2.5	3.9	1.1	versicolor	2	2
71	5.9	3.2	4.8	1.8	versicolor	2	3
72	6.1	2.8	4	1.3	versicolor	2	2
73	6.3	2.5	4.9	1.5	versicolor	2	2
74	6.1	2.8	4.7	1.2	versicolor	2	2
75	6.4	2.9	4.3	1.3	versicolor	2	2
76	6.6	3	4.4	1.4	versicolor	2	2
77	6.8	2.8	4.8	1.4	versicolor	2	2
78	6.7	3	5	1.7	versicolor	2	2
79	6	2.9	4.5	1.5	versicolor	2	2
80	5.7	2.6	3.5	1	versicolor	2	2
81	5.5	2.4	3.8	1.1	versicolor	2	2
82	5.5	2.4	3.7	1	versicolor	2	2
83	5.8	2.7	3.9	1.2	versicolor	2	2
84	6	2.7	5.1	1.6	versicolor	2	3

85	5.4	3	4.5	1.5	versicolor	2	2
86	6	3.4	4.5	1.6	versicolor	2	2
87	6.7	3.1	4.7	1.5	versicolor	2	2
88	6.3	2.3	4.4	1.3	versicolor	2	2
89	5.6	3	4.1	1.3	versicolor	2	2
90	5.5	2.5	4	1.3	versicolor	2	2
91	5.5	2.6	4.4	1.2	versicolor	2	2
92	6.1	3	4.6	1.4	versicolor	2	2
93	5.8	2.6	4	1.2	versicolor	2	2
94	5	2.3	3.3	1	versicolor	2	2
95	5.6	2.7	4.2	1.3	versicolor	2	2
96	5.7	3	4.2	1.2	versicolor	2	2
97	5.7	2.9	4.2	1.3	versicolor	2	2
98	6.2	2.9	4.3	1.3	versicolor	2	2
99	5.1	2.5	3	1.1	versicolor	2	2
100	5.7	2.8	4.1	1.3	versicolor	2	2
101	6.3	3.3	6	2.5	virginica	3	3
102	5.8	2.7	5.1	1.9	virginica	3	3
103	7.1	3	5.9	2.1	virginica	3	3
104	6.3	2.9	5.6	1.8	virginica	3	3
105	6.5	3	5.8	2.2	virginica	3	3
106	7.6	3	6.6	2.1	virginica	3	3
107	4.9	2.5	4.5	1.7	virginica	3	3
108	7.3	2.9	6.3	1.8	virginica	3	3
109	6.7	2.5	5.8	1.8	virginica	3	3
110	7.2	3.6	6.1	2.5	virginica	3	3
111	6.5	3.2	5.1	2	virginica	3	3
112	6.4	2.7	5.3	1.9	virginica	3	3
113	6.8	3	5.5	2.1	virginica	3	3
114	5.7	2.5	5	2	virginica	3	3
115	5.8	2.8	5.1	2.4	virginica	3	3
116	6.4	3.2	5.3	2.3	virginica	3	3
117	6.5	3	5.5	1.8	virginica	3	3
118	7.7	3.8	6.7	2.2	virginica	3	3
119	7.7	2.6	6.9	2.3	virginica	3	3
120	6	2.2	5	1.5	virginica	3	3
121	6.9	3.2	5.7	2.3	virginica	3	3
122	5.6	2.8	4.9	2	virginica	3	3
123	7.7	2.8	6.7	2	virginica	3	3
124	6.3	2.7	4.9	1.8	virginica	3	3
125	6.7	3.3	5.7	2.1	virginica	3	3
126	7.2	3.2	6	1.8	virginica	3	3
127	6.2	2.8	4.8	1.8	virginica	3	3
128	6.1	3	4.9	1.8	virginica	3	3
129	6.4	2.8	5.6	2.1	virginica	3	3
130	7.2	3	5.8	1.6	virginica	3	3

131	7.4	2.8	6.1	1.9	virginica	3	3
132	7.9	3.8	6.4	2	virginica	3	3
133	6.4	2.8	5.6	2.2	virginica	3	3
134	6.3	2.8	5.1	1.5	virginica	3	2
135	6.1	2.6	5.6	1.4	virginica	3	3
136	7.7	3	6.1	2.3	virginica	3	3
137	6.3	3.4	5.6	2.4	virginica	3	3
138	6.4	3.1	5.5	1.8	virginica	3	3
139	6	3	4.8	1.8	virginica	3	3
140	6.9	3.1	5.4	2.1	virginica	3	3
141	6.7	3.1	5.6	2.4	virginica	3	3
142	6.9	3.1	5.1	2.3	virginica	3	3
143	5.8	2.7	5.1	1.9	virginica	3	3
144	6.8	3.2	5.9	2.3	virginica	3	3
145	6.7	3.3	5.7	2.5	virginica	3	3
146	6.7	3	5.2	2.3	virginica	3	3
147	6.3	2.5	5	1.9	virginica	3	3
148	6.5	3	5.2	2	virginica	3	3
149	6.2	3.4	5.4	2.3	virginica	3	3
150	5.9	3	5.1	1.8	virginica	3	3

In this data there were 3 mispredictions.

3. Construct the classification matrix and report the percentage of correct classification.

Observed		Predicted			% correct
		1	2	3	
	1	50	0	0	100%
	2	0	48	2	96%
	3	0	1	49	98%

The classifiation table has been given above and percentage of correctclassification for flower type 1, 2, 3 are 100%, 96% and 98% respectively**.**

Case 2: Consider the dataset Flower Species.xlsx dataset which has data on Sepal Length, Sepal Width, Petal Length and Petal Width and Species Type for 150 different flowers and perform the following objectives.

1.Built a multinomial logistic regression model for the given problem.

Let us define our hypothesis;
Ho: All parameter of the ith class is zero.
H1: Atleast one of the parameter of the ith class is non zero.

Here we have performed the likelihood ratio test.

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests

	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	21.680	9.781	2	.008
sepal.length	13.266	1.367	2	.505
sepal.width	15.492	3.594	2	.166
petal.length	25.902	14.003	2	.001
petal.width	23.772	11.873	2	.003

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

From the table we can see that we will reject Ho in the case of petal_length and petal_width. Hence atleast one of the parameter of the class petal.length and petal.width is non zero.

Parameter Estimates

Type[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
1	Intercept	33.164	1.852E5	.000	1	1.000			
	sepal.length	11.864	5.456E4	.000	1	1.000	1.421E5	.000	.[b]
	sepal.width	13.276	2.597E4	.000	1	1.000	5.832E5	.000	.[b]
	petal.length	-26.896	2.548E4	.000	1	.999	2.086E-12	.000	.[b]
	petal.width	-38.067	.000	.	1	.	2.935E-17	2.935E-17	2.935E-17
2	Intercept	42.638	25.708	2.751	1	.097			
	sepal.length	2.465	2.394	1.060	1	.303	11.766	.108	1284.293
	sepal.width	6.681	4.480	2.224	1	.136	797.026	.123	5181847.251
	petal.length	-9.429	4.737	3.962	1	.047	8.033E-5	7.457E-9	.865
	petal.width	-18.286	9.743	3.523	1	.061	1.144E-8	5.828E-17	2.246

a. The reference category is: 3.

b. Floating point overflow occurred while computing this statistic. Its value is therefore set to system missing.

Here we have two tables as SPSS uses flower type 3 as a reference and then compute the parameter estimates for flower type 1 and flower type 2.

2. Obtain the predicted class for each flower using the multinomial logistic regression.

Flower No.	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	y	Predicted group
1	5.1	3.5	1.4	0.2	setosa	1	1
2	4.9	3	1.4	0.2	setosa	1	1
3	4.7	3.2	1.3	0.2	setosa	1	1

4	4.6	3.1	1.5	0.2	setosa	1	1
5	5	3.6	1.4	0.2	setosa	1	1
6	5.4	3.9	1.7	0.4	setosa	1	1
7	4.6	3.4	1.4	0.3	setosa	1	1
8	5	3.4	1.5	0.2	setosa	1	1
9	4.4	2.9	1.4	0.2	setosa	1	1
10	4.9	3.1	1.5	0.1	setosa	1	1
11	5.4	3.7	1.5	0.2	setosa	1	1
12	4.8	3.4	1.6	0.2	setosa	1	1
13	4.8	3	1.4	0.1	setosa	1	1
14	4.3	3	1.1	0.1	setosa	1	1
15	5.8	4	1.2	0.2	setosa	1	1
16	5.7	4.4	1.5	0.4	setosa	1	1
17	5.4	3.9	1.3	0.4	setosa	1	1
18	5.1	3.5	1.4	0.3	setosa	1	1
19	5.7	3.8	1.7	0.3	setosa	1	1
20	5.1	3.8	1.5	0.3	setosa	1	1
21	5.4	3.4	1.7	0.2	setosa	1	1
22	5.1	3.7	1.5	0.4	setosa	1	1
23	4.6	3.6	1	0.2	setosa	1	1
24	5.1	3.3	1.7	0.5	setosa	1	1
25	4.8	3.4	1.9	0.2	setosa	1	1
26	5	3	1.6	0.2	setosa	1	1
27	5	3.4	1.6	0.4	setosa	1	1
28	5.2	3.5	1.5	0.2	setosa	1	1
29	5.2	3.4	1.4	0.2	setosa	1	1
30	4.7	3.2	1.6	0.2	setosa	1	1
31	4.8	3.1	1.6	0.2	setosa	1	1
32	5.4	3.4	1.5	0.4	setosa	1	1
33	5.2	4.1	1.5	0.1	setosa	1	1
34	5.5	4.2	1.4	0.2	setosa	1	1
35	4.9	3.1	1.5	0.2	setosa	1	1
36	5	3.2	1.2	0.2	setosa	1	1
37	5.5	3.5	1.3	0.2	setosa	1	1
38	4.9	3.6	1.4	0.1	setosa	1	1
39	4.4	3	1.3	0.2	setosa	1	1
40	5.1	3.4	1.5	0.2	setosa	1	1
41	5	3.5	1.3	0.3	setosa	1	1
42	4.5	2.3	1.3	0.3	setosa	1	1
43	4.4	3.2	1.3	0.2	setosa	1	1
44	5	3.5	1.6	0.6	setosa	1	1
45	5.1	3.8	1.9	0.4	setosa	1	1
46	4.8	3	1.4	0.3	setosa	1	1
47	5.1	3.8	1.6	0.2	setosa	1	1
48	4.6	3.2	1.4	0.2	setosa	1	1
49	5.3	3.7	1.5	0.2	setosa	1	1

50	5	3.3	1.4	0.2	setosa	1	1
51	7	3.2	4.7	1.4	versicolor	2	2
52	6.4	3.2	4.5	1.5	versicolor	2	2
53	6.9	3.1	4.9	1.5	versicolor	2	2
54	5.5	2.3	4	1.3	versicolor	2	2
55	6.5	2.8	4.6	1.5	versicolor	2	2
56	5.7	2.8	4.5	1.3	versicolor	2	2
57	6.3	3.3	4.7	1.6	versicolor	2	2
58	4.9	2.4	3.3	1	versicolor	2	2
59	6.6	2.9	4.6	1.3	versicolor	2	2
60	5.2	2.7	3.9	1.4	versicolor	2	2
61	5	2	3.5	1	versicolor	2	2
62	5.9	3	4.2	1.5	versicolor	2	2
63	6	2.2	4	1	versicolor	2	2
64	6.1	2.9	4.7	1.4	versicolor	2	2
65	5.6	2.9	3.6	1.3	versicolor	2	2
66	6.7	3.1	4.4	1.4	versicolor	2	2
67	5.6	3	4.5	1.5	versicolor	2	2
68	5.8	2.7	4.1	1	versicolor	2	2
69	6.2	2.2	4.5	1.5	versicolor	2	2
70	5.6	2.5	3.9	1.1	versicolor	2	2
71	5.9	3.2	4.8	1.8	versicolor	2	2
72	6.1	2.8	4	1.3	versicolor	2	2
73	6.3	2.5	4.9	1.5	versicolor	2	2
74	6.1	2.8	4.7	1.2	versicolor	2	2
75	6.4	2.9	4.3	1.3	versicolor	2	2
76	6.6	3	4.4	1.4	versicolor	2	2
77	6.8	2.8	4.8	1.4	versicolor	2	2
78	6.7	3	5	1.7	versicolor	2	2
79	6	2.9	4.5	1.5	versicolor	2	2
80	5.7	2.6	3.5	1	versicolor	2	2
81	5.5	2.4	3.8	1.1	versicolor	2	2
82	5.5	2.4	3.7	1	versicolor	2	2
83	5.8	2.7	3.9	1.2	versicolor	2	2
84	6	2.7	5.1	1.6	versicolor	2	3
85	5.4	3	4.5	1.5	versicolor	2	2
86	6	3.4	4.5	1.6	versicolor	2	2
87	6.7	3.1	4.7	1.5	versicolor	2	2
88	6.3	2.3	4.4	1.3	versicolor	2	2
89	5.6	3	4.1	1.3	versicolor	2	2
90	5.5	2.5	4	1.3	versicolor	2	2
91	5.5	2.6	4.4	1.2	versicolor	2	2
92	6.1	3	4.6	1.4	versicolor	2	2
93	5.8	2.6	4	1.2	versicolor	2	2
94	5	2.3	3.3	1	versicolor	2	2
95	5.6	2.7	4.2	1.3	versicolor	2	2

96	5.7	3	4.2	1.2	versicolor	2	2
97	5.7	2.9	4.2	1.3	versicolor	2	2
98	6.2	2.9	4.3	1.3	versicolor	2	2
99	5.1	2.5	3	1.1	versicolor	2	2
100	5.7	2.8	4.1	1.3	versicolor	2	2
101	6.3	3.3	6	2.5	virginica	3	3
102	5.8	2.7	5.1	1.9	virginica	3	3
103	7.1	3	5.9	2.1	virginica	3	3
104	6.3	2.9	5.6	1.8	virginica	3	3
105	6.5	3	5.8	2.2	virginica	3	3
106	7.6	3	6.6	2.1	virginica	3	3
107	4.9	2.5	4.5	1.7	virginica	3	3
108	7.3	2.9	6.3	1.8	virginica	3	3
109	6.7	2.5	5.8	1.8	virginica	3	3
110	7.2	3.6	6.1	2.5	virginica	3	3
111	6.5	3.2	5.1	2	virginica	3	3
112	6.4	2.7	5.3	1.9	virginica	3	3
113	6.8	3	5.5	2.1	virginica	3	3
114	5.7	2.5	5	2	virginica	3	3
115	5.8	2.8	5.1	2.4	virginica	3	3
116	6.4	3.2	5.3	2.3	virginica	3	3
117	6.5	3	5.5	1.8	virginica	3	3
118	7.7	3.8	6.7	2.2	virginica	3	3
119	7.7	2.6	6.9	2.3	virginica	3	3
120	6	2.2	5	1.5	virginica	3	3
121	6.9	3.2	5.7	2.3	virginica	3	3
122	5.6	2.8	4.9	2	virginica	3	3
123	7.7	2.8	6.7	2	virginica	3	3
124	6.3	2.7	4.9	1.8	virginica	3	3
125	6.7	3.3	5.7	2.1	virginica	3	3
126	7.2	3.2	6	1.8	virginica	3	3
127	6.2	2.8	4.8	1.8	virginica	3	3
128	6.1	3	4.9	1.8	virginica	3	3
129	6.4	2.8	5.6	2.1	virginica	3	3
130	7.2	3	5.8	1.6	virginica	3	3
131	7.4	2.8	6.1	1.9	virginica	3	3
132	7.9	3.8	6.4	2	virginica	3	3
133	6.4	2.8	5.6	2.2	virginica	3	3
134	6.3	2.8	5.1	1.5	virginica	3	2
135	6.1	2.6	5.6	1.4	virginica	3	3
136	7.7	3	6.1	2.3	virginica	3	3
137	6.3	3.4	5.6	2.4	virginica	3	3
138	6.4	3.1	5.5	1.8	virginica	3	3
139	6	3	4.8	1.8	virginica	3	3
140	6.9	3.1	5.4	2.1	virginica	3	3
141	6.7	3.1	5.6	2.4	virginica	3	3

142	6.9	3.1	5.1	2.3	virginica	3	3
143	5.8	2.7	5.1	1.9	virginica	3	3
144	6.8	3.2	5.9	2.3	virginica	3	3
145	6.7	3.3	5.7	2.5	virginica	3	3
146	6.7	3	5.2	2.3	virginica	3	3
147	6.3	2.5	5	1.9	virginica	3	3
148	6.5	3	5.2	2	virginica	3	3
149	6.2	3.4	5.4	2.3	virginica	3	3
150	5.9	3	5.1	1.8	virginica	3	3

4. Construct the classification Matrix and report the percentage of correct classification.

The classifiation table has been given below and percentage of correct classification for flower types: sentosa(1),versicolor(2),virginica (3) are 100%, 98% and 98% respectively.

Classification

	Predicted			
Observed	1	2	3	Percent Correct
1	50	0	0	100.0%
2	0	49	1	98.0%
3	0	1	49	98.0%
Overall Percentage	33.3%	33.3%	33.3%	98.7%

COMPARATIVE STUDY OF THE TWO APPROACHES OF MULTICLASS CLASSIFICATION FOR FLOWER SPECIES PREDICTION PROBLEM

In the binary logistic approach, 3 observations were classified incorrectly, whereas in multinomial logistic approach, only 2 were classified wrongly.

The classification matrices are-

- Binary Logistic

Observed		Predicted			% correct
		1	2	3	
	1	50	0	0	100%
	2	0	48	2	96%
	3	0	1	49	98%

- Multinomial Logistic-

Classification

Observed	Predicted			
	1	2	3	Percent Correct
1	50	0	0	100.0%
2	0	49	1	98.0%
3	0	1	49	98.0%
Overall Percentage	33.3%	33.3%	33.3%	98.7%

Both the methods classify the Group 1 (Species Setosa) perfectly.

For the Group 2 (Versicolor), multinomial logistic regression classifies 1 one more observation correctly than binary logistic regression.

For binary logistic, the overall percentage of correct classification is 98%, whereas for multinomial logistic method, it is 98.7%.