Pima Indian Diabetes

1. Fit a Linear Discriminant Analysis model (function lda in package MASS) and calculate the misclassification error on the provided test set. (2 marks)

```
library(MASS)
pima <- read.csv("pima.csv")
pima_test <- read.csv("pima_test.csv")
pima_lda <- lda(pima$type == 1 ~ ., data=pima, prior=c(0.66, 0.34))
pred_lda <- predict(pima_lda, newdata=pima_test)
lda_matrix <- table(pima_test$type, pred_lda$class)
lda_matrix
```

```
##
##     FALSE TRUE
##   0   198   25
##   1    42   67
```

```
(25+42)/(25+42+198+67)
```

```
## [1] 0.2018072
```

The misclassification error is approximately 0.20

2. Which kind of misclassification is more common in the test data: patients with diabetes misclassified as healthy, or healthy patients misclassified as having diabetes? (2 marks)

```
25/(25+42+198+67)
```

```
## [1] 0.0753012
```

```
42/(25+42+198+67)
```

```
## [1] 0.126506
```

The 'patient with diabetes misclassified as healthy' has a higher proportion in the misclassification error, hence this misclassification is more common in the test data.

3. Fit a Quadratic Disciminant Analysis model (function qda in package MASS). Write down the misclassification error (Question 1) for the QDA model. (1 marks)

```
pima_qda <- qda(pima$type == 1 ~ ., data=pima, prior=c(0.66, 0.34))
pred_qda <- predict(pima_qda, newdata=pima_test)
qda_matrix <- table(pima_test$type, pred_qda$class)
qda_matrix
```

```
##
##     FALSE TRUE
##   0   194   29
##   1    47   62
```

```
(29+47)/(194+29+47+62)
```

```
## [1] 0.2289157
```

The misclassification error is approximately 0.23.

4. A health organisation wants you to recommend one of the two models for diagnosing diabetes. What would you tell them? Explain your decision in a way that a non-statistician could understand. (2 marks)

I would recommend the LDA model as its misclassification error (0.20) is lower than that of the QDA model (0.23). This means that the LDA model is more likely to classify healthy patients as healthy and diabetic patients as diabetic.

5. Fit a logistic regression model. What is the test error for this model? (1 mark)

```
lrm <- glm(pima$type == 1 ~ ., data = pima, family=binomial())
tst_pred_lr <- ifelse(predict(lrm, newdata = pima_test,
                              type="response") > 0.5, 1, 0)
tst_tab_lr<- table(pima_test$type, tst_pred_lr)
tst_tab_lr
```

```
##    tst_pred_lr
##       0   1
##   0 200  23
##   1  43  66
```

```
(23+43)/(200+43+23+66)
```

```
## [1] 0.1987952
```

The test error is approximately 0.20.

6. A woman wants to know about her diabetes status. Following data is available:

npreg glu bp skin bmi ped age

5 111 81 33 25.1 0.36 48

Predict the diabetes status by logistic regression without using the R function predict. (And outline the calculations involved) (2 marks)

```
summary(lrm)
```

```
##
## Call:
## glm(formula = pima$type == 1 ~ ., family = binomial(), data = pima)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.773062   1.770386  -5.520 3.38e-08 ***
## npreg        0.103183   0.064694   1.595  0.11073
## glu          0.032117   0.006787   4.732 2.22e-06 ***
## bp          -0.004768   0.018541  -0.257  0.79707
## skin        -0.001917   0.022500  -0.085  0.93211
## bmi          0.083624   0.042827   1.953  0.05087 .
## ped          1.820410   0.665514   2.735  0.00623 **
```

```
## age          0.041184   0.022091   1.864  0.06228 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 256.41  on 199  degrees of freedom
## Residual deviance: 178.39  on 192  degrees of freedom
## AIC: 194.39
##
## Number of Fisher Scoring iterations: 5
```

```
numerator = exp(-9.773062 + 0.103183*5 + 0.032117*111 + -0.004768*81 +
                  -0.001917*33 + 0.083624*25.1 + 1.820410*0.36 + 0.041184*48)
denominator = (1 + exp(-9.773062 + 0.103183*5 + 0.032117*111 + -0.004768*81 +
                         -0.001917*33 + 0.083624*25.1 +
                         1.820410*0.36 + 0.041184*48))
prob_of_having_diabetes = numerator/denominator
prob_of_having_diabetes
```

```
## [1] 0.1961573
```

By using logistic regression, the probability of this woman having diabetes is approximately 20%. This woman does not belong in the diabetic group as she's less than 0.5.