Critical Thinking Group 4: DATA621 Homework 2

Table of Contents

[TEAM Members: 1](#_Toc34027333)

[1 Overview 2](#_Toc34027334)

[2 Deliverables 2](#_Toc34027335)

[3 Task 1: Download Data Set 2](#_Toc34027336)

[4 Task 2: Confusion Matrix 3](#_Toc34027337)

[5 Task 3: Accuracy 5](#_Toc34027338)

[6 Task 4: Classification Error Rate 6](#_Toc34027339)

[7 Task 5: Precision 7](#_Toc34027340)

[8 Task 6: Sensitivity 8](#_Toc34027341)

[9 Task 7: Specificity 8](#_Toc34027342)

[10 Task 8: F1 Score 9](#_Toc34027343)

[11 Task 9: Prove 0 < F1Score < 1 9](#_Toc34027344)

[12 Task 10: ROC Curve 10](#_Toc34027345)

[13 Task 11: Produce All Metrics 12](#_Toc34027346)

[14 Task 12: Package: Caret 13](#_Toc34027347)

[15 Task 13: Package: pROC 14](#_Toc34027348)

[Appendix 15](#_Toc34027349)

## TEAM Members:

Rajwant Mishra  
Priya Shaji  
Debabrata Kabiraj  
Isabel Ramesar  
Sin Ying Wong  
Fan Xu

-------------------------------------------------

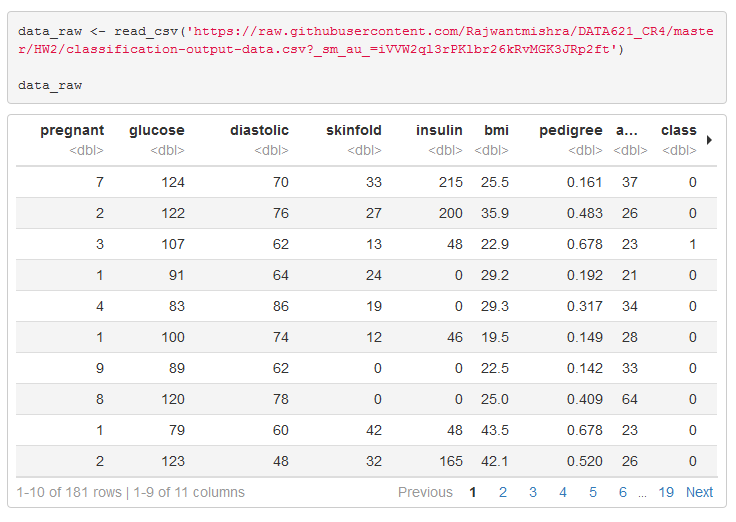
**1 Overview**

In this homework assignment, you will work through various classification metrics. You will be asked to create functions in R to carry out the various calculations. You will also investigate some functions in packages that will let you obtain the equivalent results. Finally, you will create graphical output that also can be used to evaluate the output of classification models, such as binary logistic regression.

**2 Deliverables**

Upon following the instructions below, use your created R functions and the other packages to generate the classification metrics for the provided data set. A write-up of your solutions submitted in PDF format.

**3 Task 1: Download Data Set**



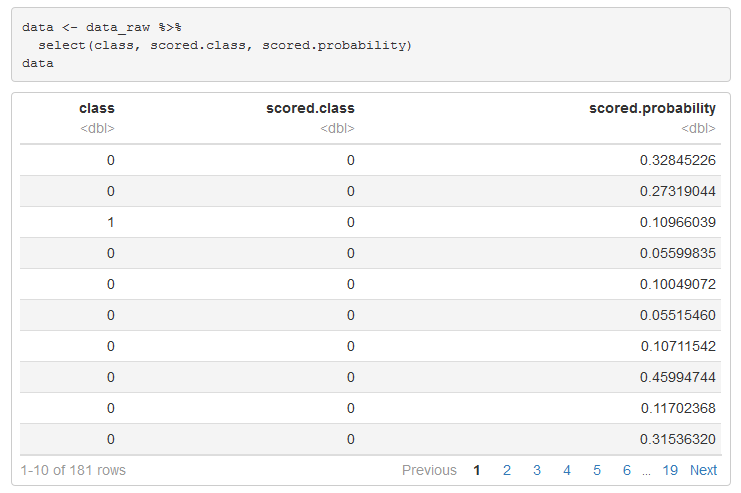
**4 Task 2: Confusion Matrix**

The data set has three key columns we will use:

class: the actual class for the observation

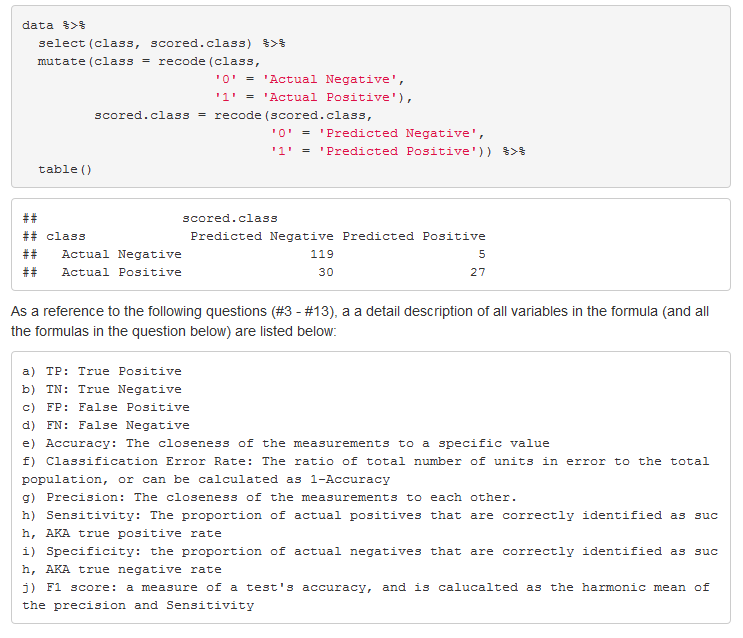
scored.class: the predicted class for the observation (based on a threshold of 0.5)

scored.probability: the predicted probability of success for the observation



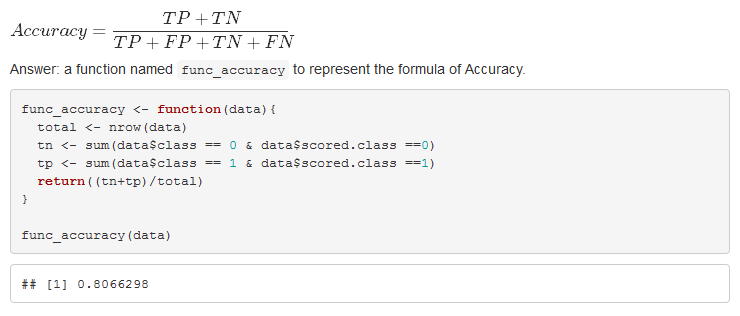
Use the table() function to get the raw confusion matrix for this scored dataset. Make sure you understand the output. In particular, do the rows represent the actual or predicted class? The columns?

Answer: the field class (the rows) represent the actual class, and the field scored.class (the columns) represent the predicted class.



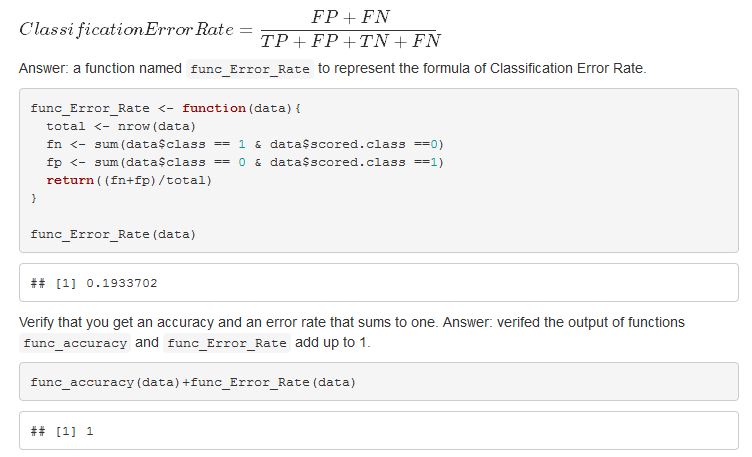
**5 Task 3: Accuracy**

Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the accuracy of the predictions.



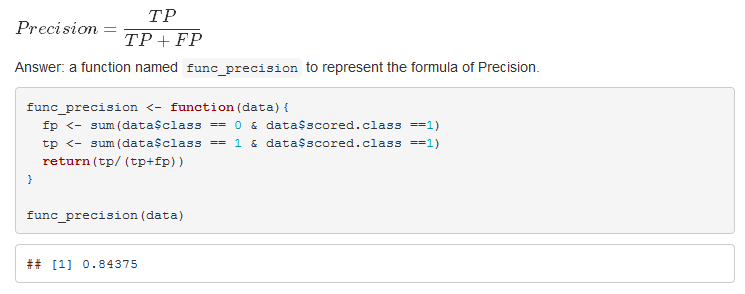
**6 Task 4: Classification Error Rate**

Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the classification error rate of the predictions.



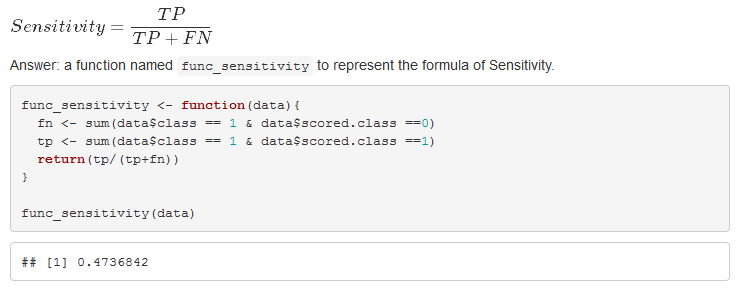
**7 Task 5: Precision**

Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the precision of the predictions.



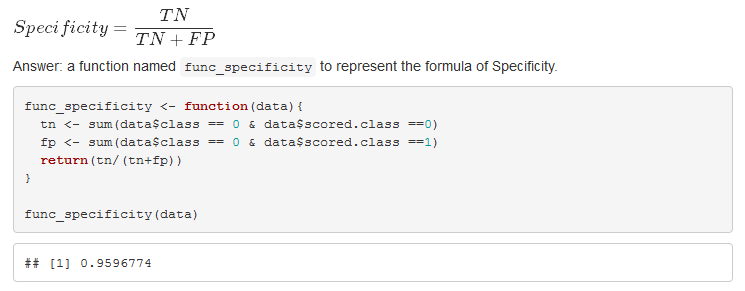
**8 Task 6: Sensitivity**

Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the sensitivity of the predictions. Sensitivity is also known as recall.



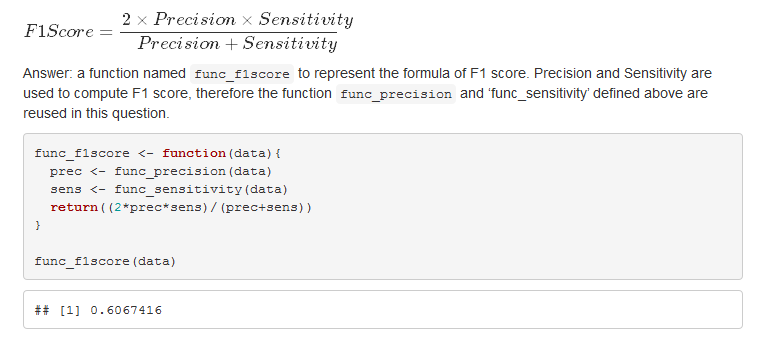
**9 Task 7: Specificity**

Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the specificity of the predictions.



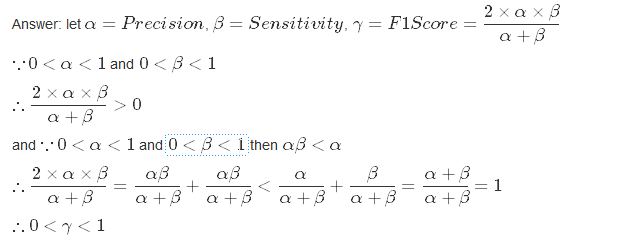
**10 Task 8: F1 Score**

Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the F1 score of the predictions.



**11 Task 9: Prove 0 < F1Score < 1**

Before we move on, let’s consider a question that was asked: What are the bounds on the F1 score? Show that the F1 score will always be between 0 and 1. (Hint: If 0<a<1 and 0<b<1 then ab<a.)



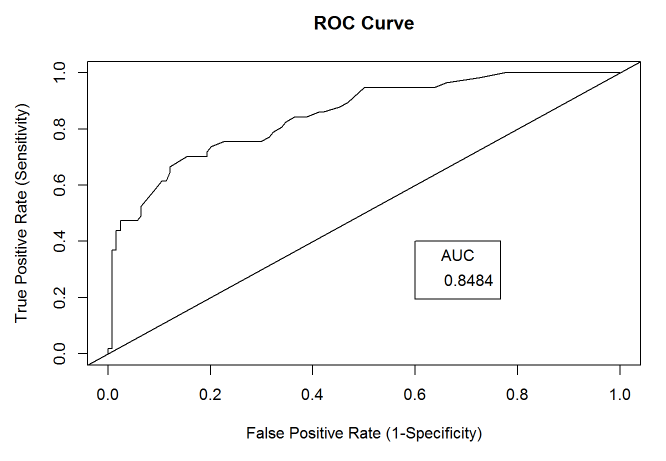
# 12 Task 10: ROC Curve

Write a function that generates an ROC curve from a data set with a true classification column (class in our example) and a probability column (scored.probability in our example). Your function should return a list that includes the plot of the ROC curve and a vector that contains the calculated area under the curve (AUC). Note that I recommend using a sequence of thresholds ranging from 0 to 1 at 0.01 intervals.

Answer: ROC curve (short form of Receiver Operating Characteristic curve), is a graphical phot that illustrates the diagonostic ability of a binary classifier system as its discrimination threshold is varied (Reference: Wikipedia).

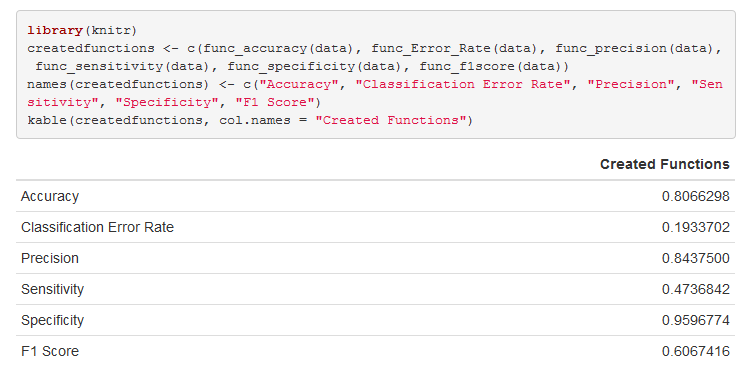
The ROC curve is created by plotting the true positive rate (TPR, or a.k.a Senstivity) against the false positive rate (FPR, can be calculated as (1-Specificity)) at various threshold settings.





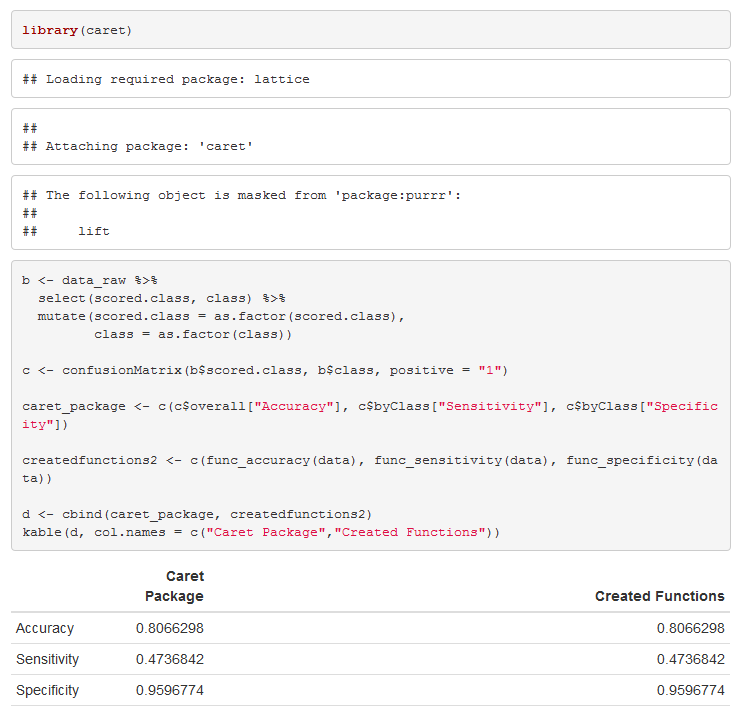
# 13 Task 11: Produce All Metrics

Use your created R functions and the provided classification output data set to produce all of the classification metrics discussed above.



# 14 Task 12: Package: Caret

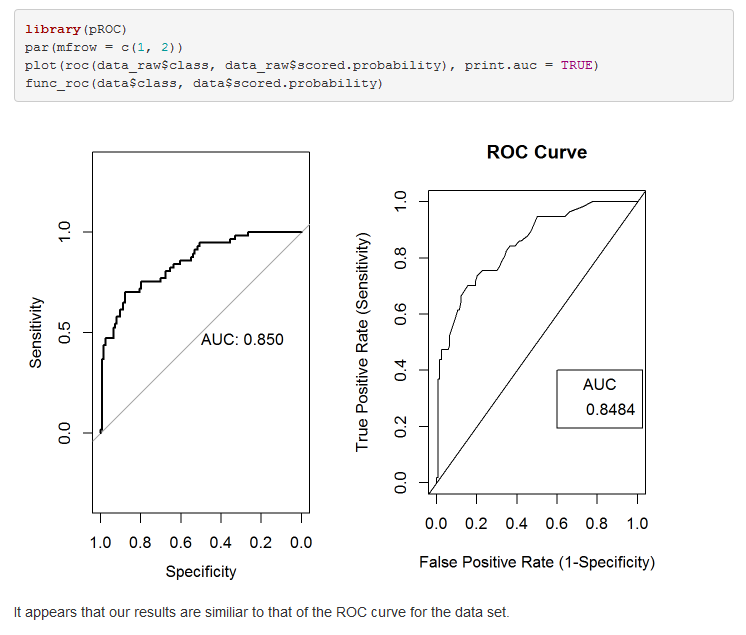
Investigate the caret package. In particular, consider the functions confusionMatrix, sensitivity, and specificity. Apply the functions to the data set. How do the results compare with your own functions?



The results from the Caret package and the functions confusionMatrix, sensitivity, and specificity are the same.

# 15 Task 13: Package: pROC

Investigate the pROC package. Use it to generate an ROC curve for the data set. How do the results compare with your own functions?



# Appendix

<https://github.com/Rajwantmishra/DATA621_CR4/blob/master/HW2/DATA%20621%20Homework%20%232%20-%20Ver%203.Rmd>

**Thank you**