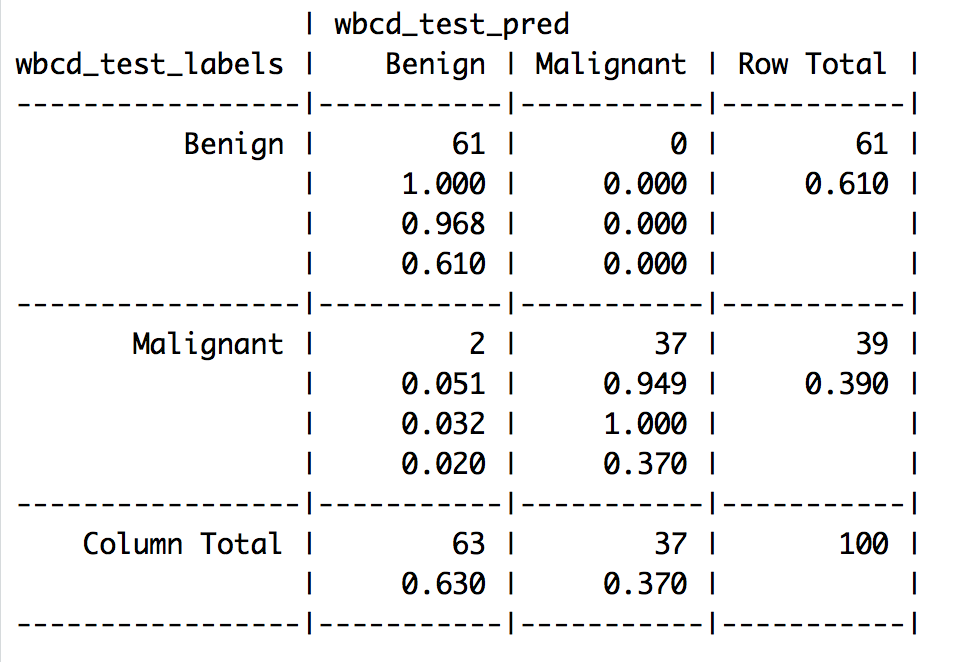
Statistics 6620

Homework 2

Surabhi Asati

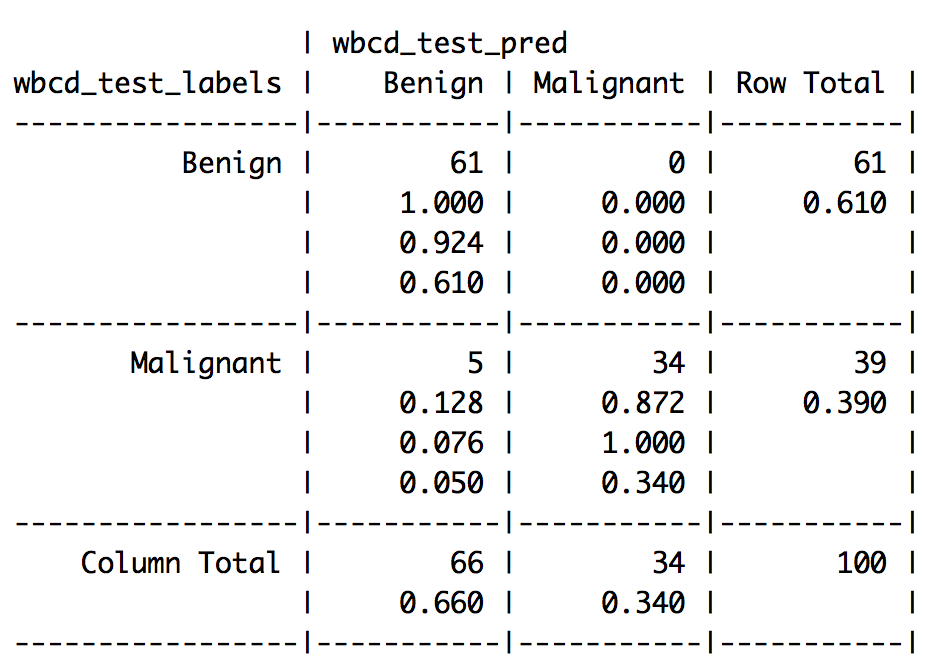
1. Perform the cancer diagnosis kNN analysis. Produce a report explaining the data, the analysis, and the findings.
2. Show the prediction that the algorithm produced.

cross tabulation of predicted vs. actual for normal datasets:



Correctly classified = 98

cross tabulation of predicted vs. actual for z-score standardize datasets:

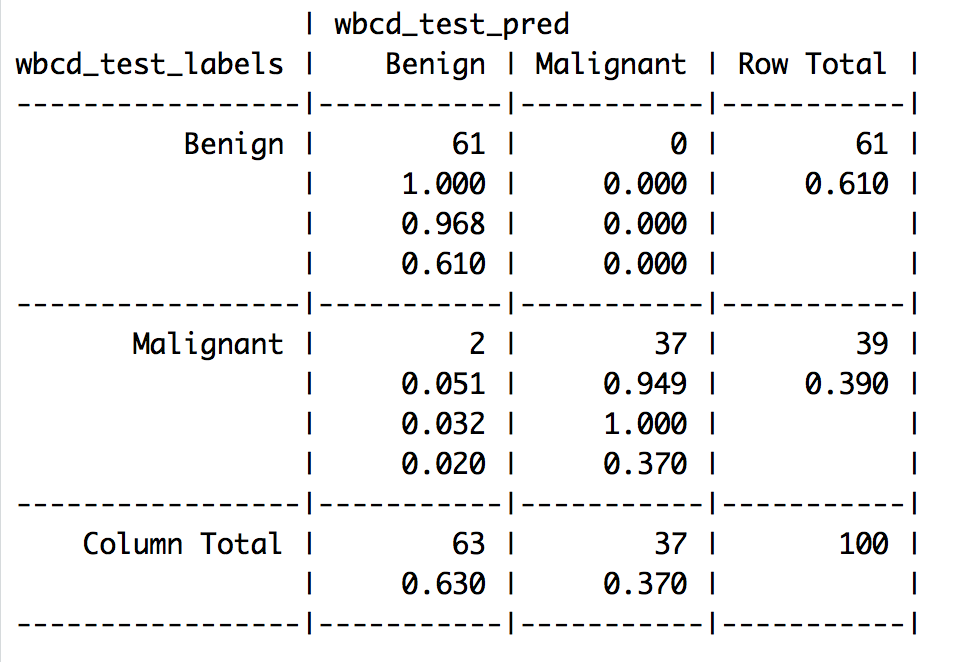


Correctly classified = 95

Predictions are better on normalizing the dataset rather Z-score scaling.

1. Give the Accuracy of the predictions. See Page 318 (or 299).

With the 2\*2 Confusion Matrix, Prediction accuracy for normal datasets:



* accuracy = TP + TN

TP+TN+FP+FN

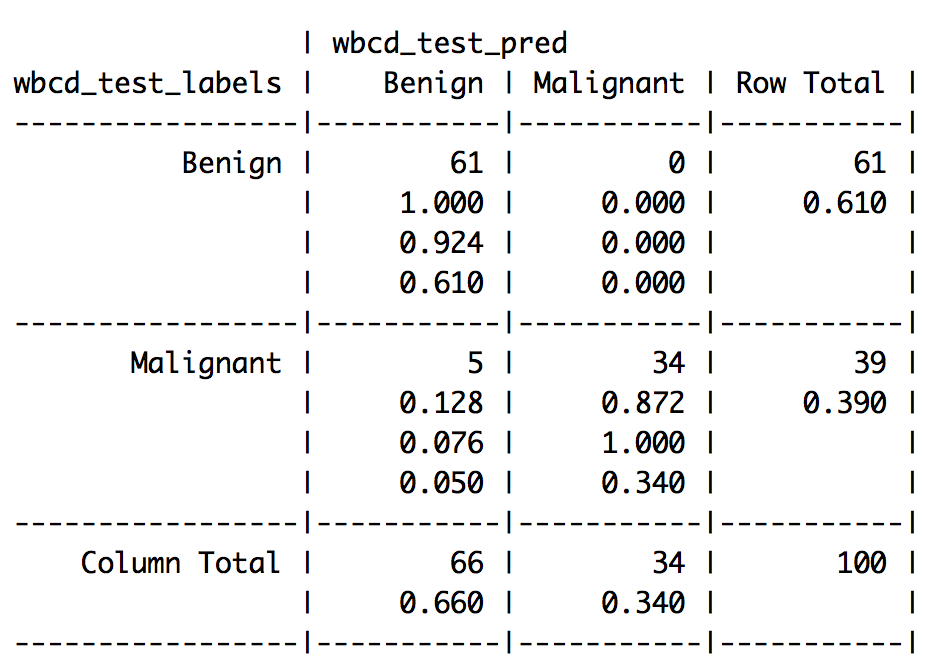


= (61+37) / (61+37+0+2)

= 98/100

= 98%

With the 2\*2 Confusion Matrix, Prediction accuracy for Z scaled datasets:



accuracy = TP + TN

TP+TN+FP+FN



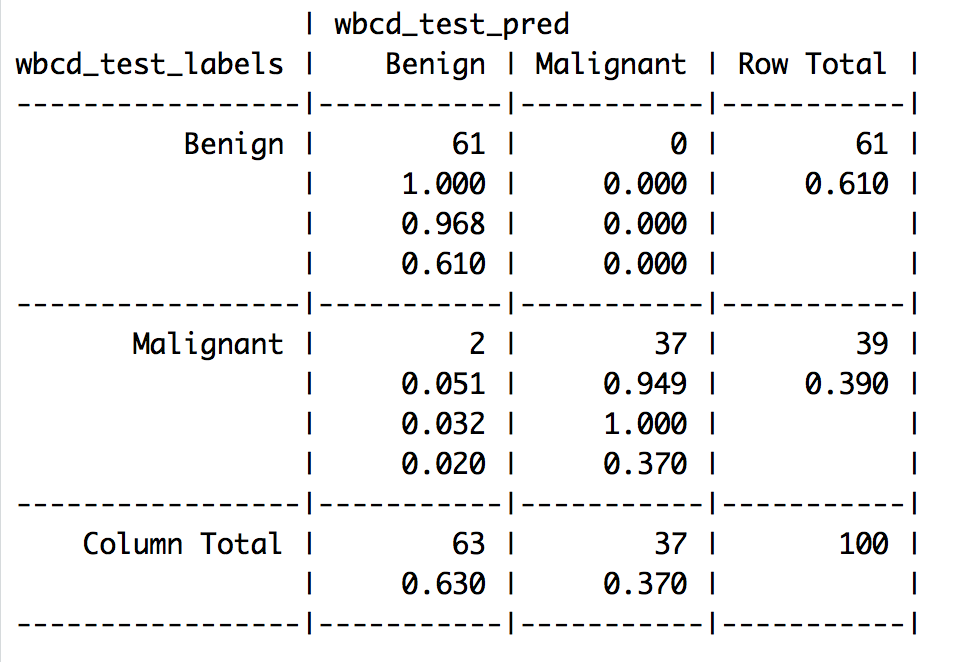
= (61+34) / (61+34+0+5)

= 95/100

= 95%

1. Include the confusion matrix.

The 2\*2 Confusion Matrix, for normal datasets:



Where,

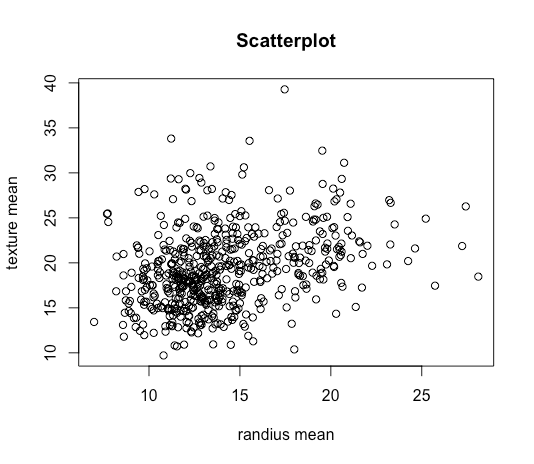
TRUE NEGATIVE = 37

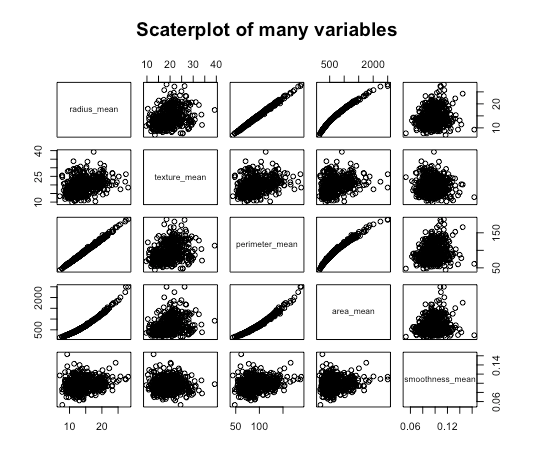
TRUE POSITIVE = 61

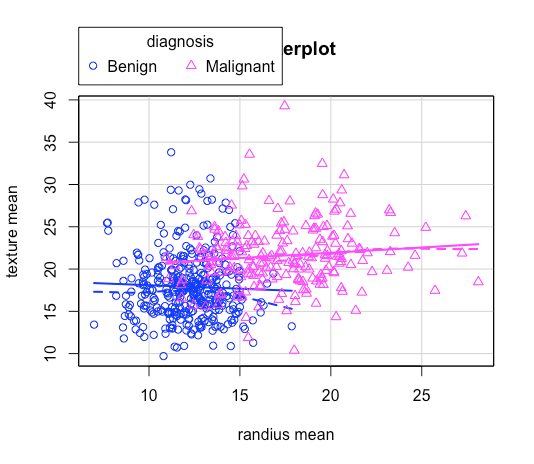
FALSE POSITIVE = 0

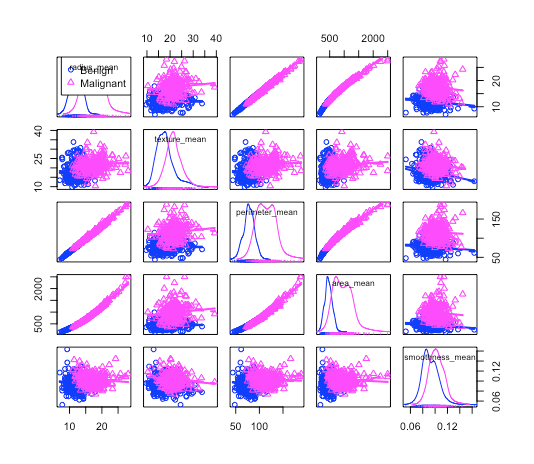
FALSE NEGATIVE = 2

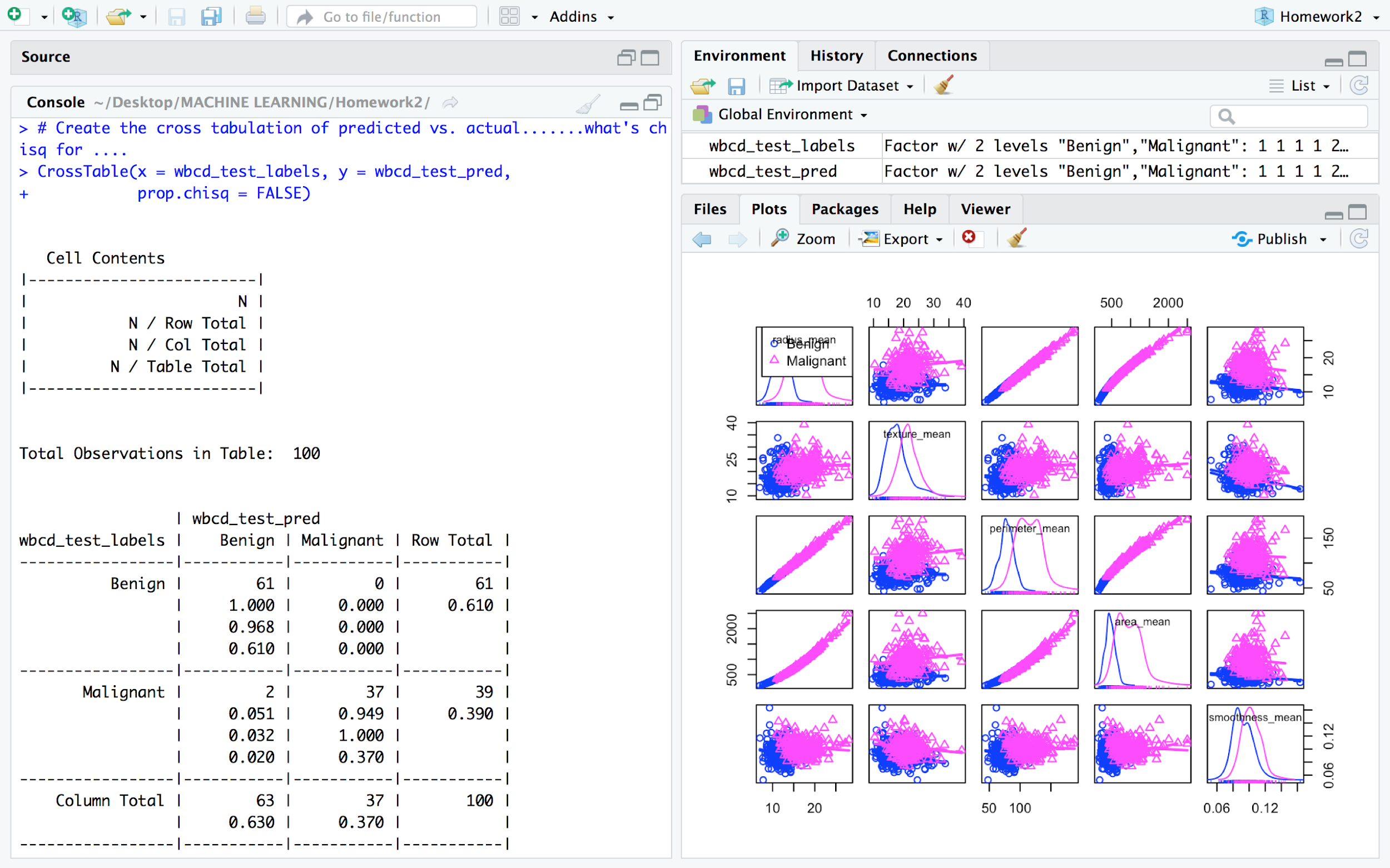
1. Graphs as follows:













1. Find an interesting dataset from the UCI ML Repository that is appropriate for applying the kNN algorithm and load the data into R and proceed to classify the data using kNN.

1. Title: Johns Hopkins University Ionosphere database  
  
 2. Source Information:  
 -- Donor: Vince Sigillito (vgs@aplcen.apl.jhu.edu)  
 -- Date: 1989

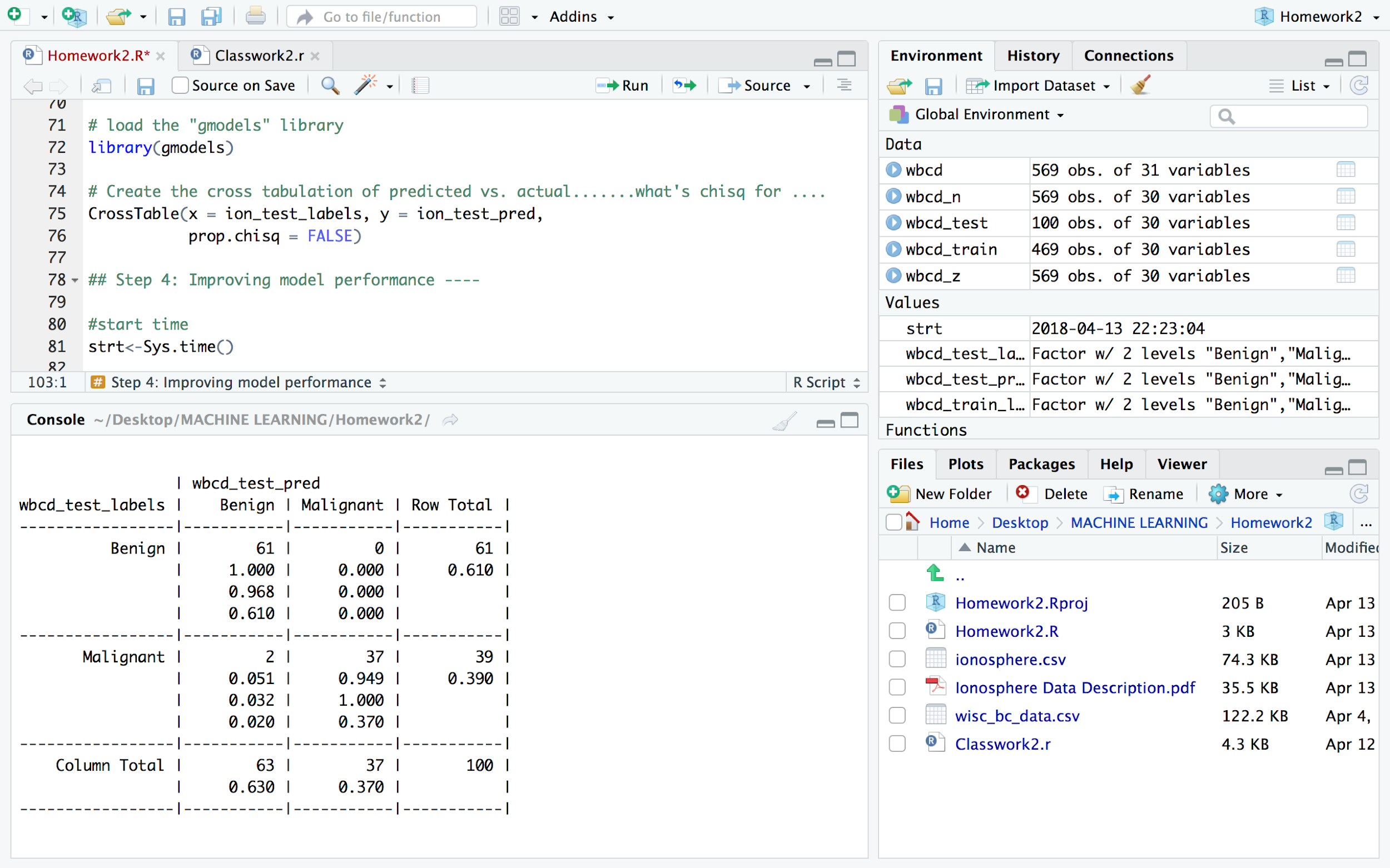
3. Relevant Information:  
 This radar data was collected by a system in Goose Bay, Labrador. The system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere.

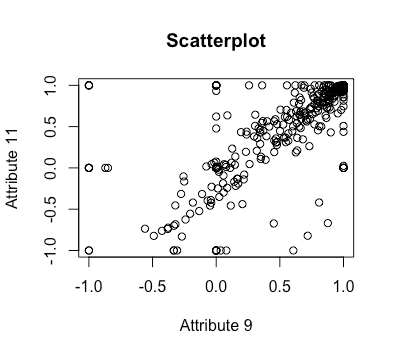
Received signals were processed using an autocorrelation function whose arguments are the time of a pulse and the pulse number. There were 17 pulse numbers for the Goose Bay system. Instances in this database are described by 2 attributes per pulse number, corresponding to the complex values returned by the function resulting from the complex electromagnetic signal.  
  
5. Number of Instances: 351  
  
6. Number of Attributes: 34 plus the class attribute  
 -- All 34 predictor attributes are continuous  
  
7. Attribute Information:   
 -- All 34 are continuous, as described above  
 -- The 35th attribute is either "good" or "bad" according to the definition  
 summarized above. This is a binary classification task.  
  
8. Missing Values: None

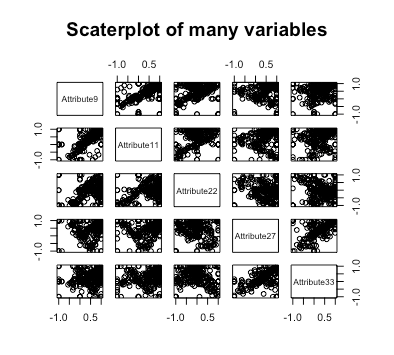
Results:

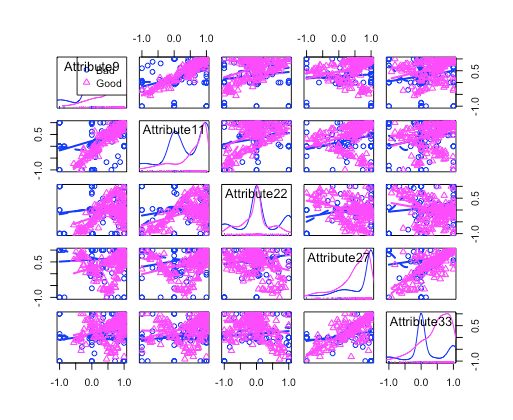
All the attributes hold value between 0 to 1. So no need to scale or normalize the data.

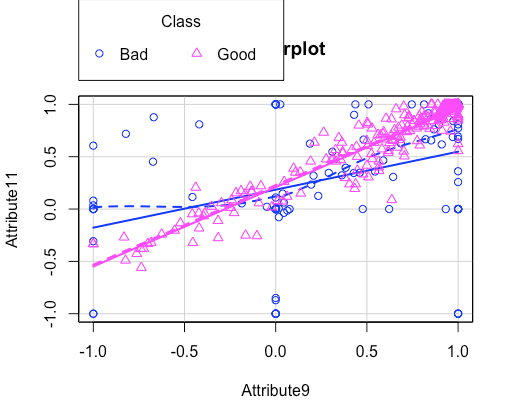
Correctly classified: 98% for K = 5, 21

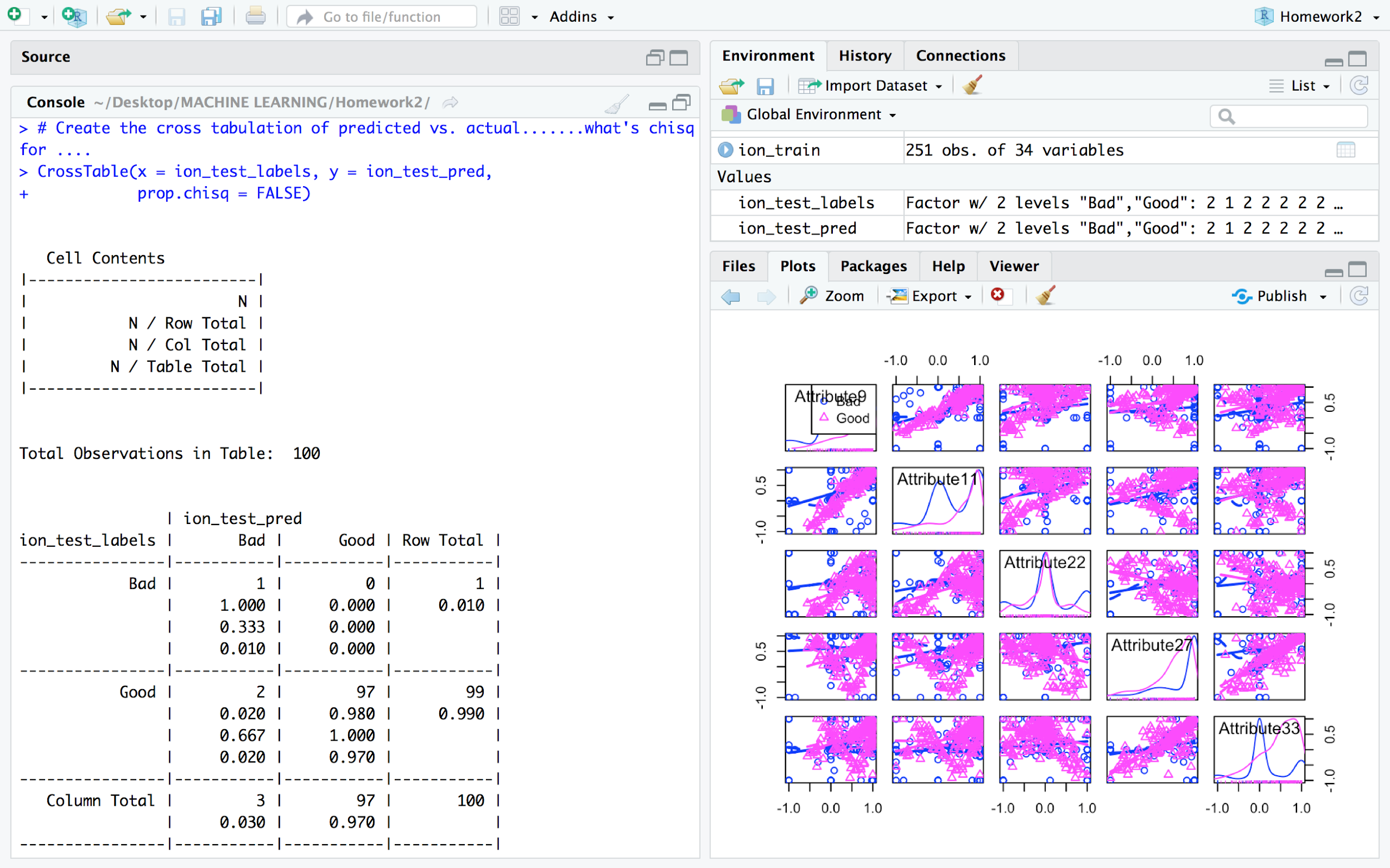


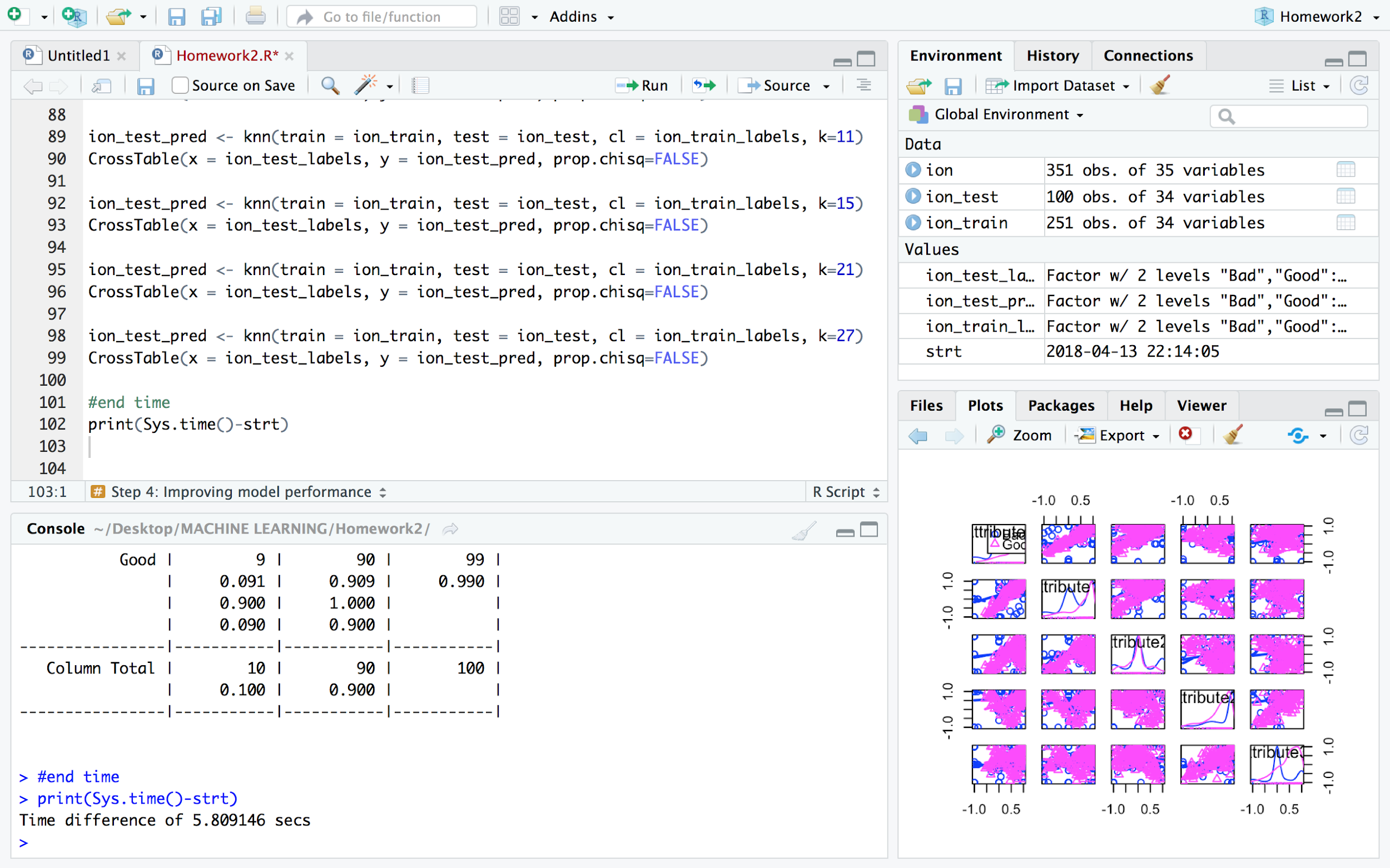












1. Do problem 7a,b,c, see page 54, in [An Introduction to Statistical Learning](http://www-bcf.usc.edu/~gareth/ISL/).

a) Euclidean distance between points a and b is the length of the line segment connecting them

1. sqrt(3^2) = 3
2. sqrt(2^2) = 2
3. sqrt(1^2 + 3^2) = sqrt(10) = 3.16 4
4. sqrt(1^2 + 2^2) = sqrt(5) = 2.24
5. sqrt(1^2 + 1^2) = sqrt(2) = 1.41
6. sqrt(1^2 + 1^2 + 1^2) = sqrt(3) = 1.73

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Observations | X1 | X2 | X3 | Y | DISTANCE |
| 1 | 0 | 3 | 0 | RED | 3 |
| 2 | 2 | 0 | 0 | RED | 2 |
| 3 | 0 | 1 | 3 | RED | 3.16 |
| 4 | 0 | 1 | 2 | GREEN | 2.24 |
| 5 | -1 | 0 | 1 | GREEN | 1.41 |
| 6 | 1 | 1 | 1 | RED | 1.73 |

b) It is the class from the nearest neighbor - the observation 5 - Green.

c) It is the average from the three nearest neighbors - observations 5, 6 and 2 -, so it is Red.