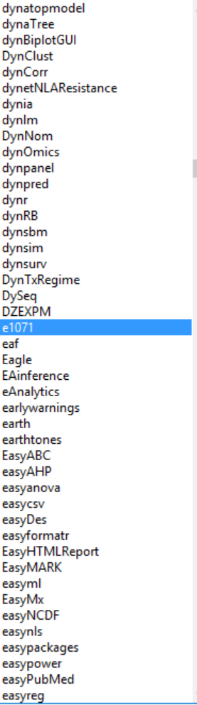
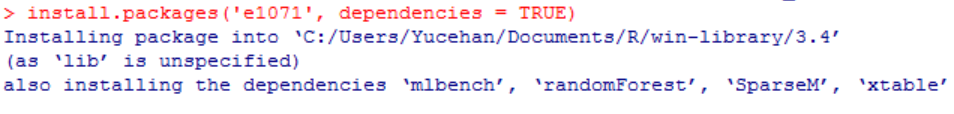
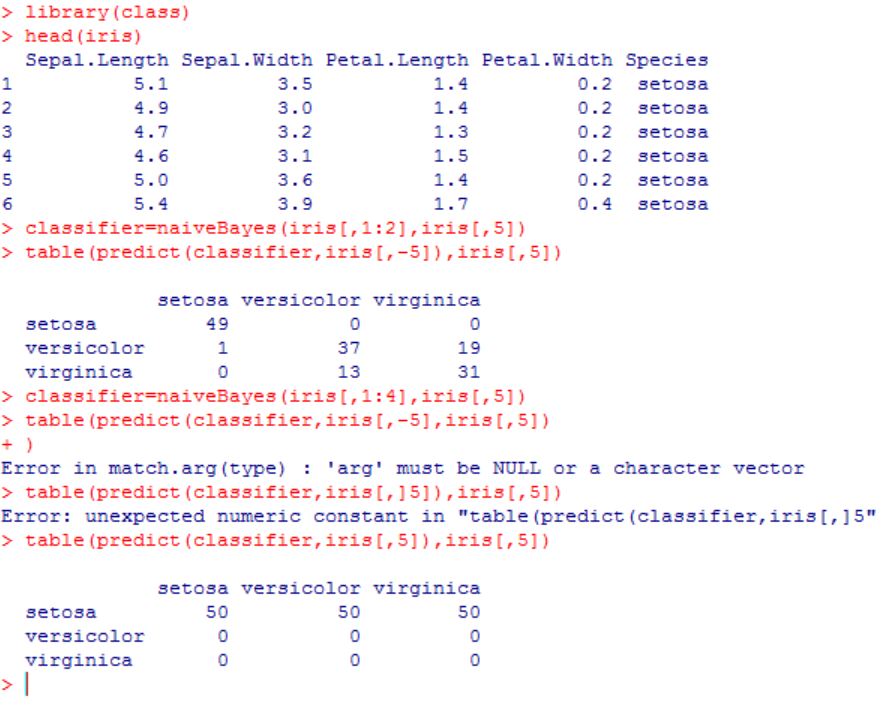
CIS/STA 3920, Lawrence Tatum

Yucehan Kucukmotor

Homework #6

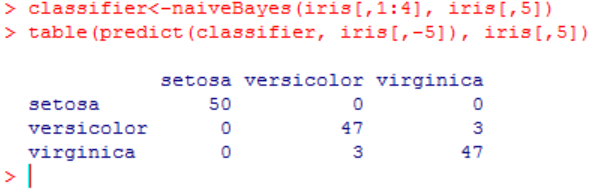
6.1)

Called library(class) and then ran the naiveBayes function. Tabled it out using the table function.

4 variables we have are correlated, and is helpful for a good prediction of the species.

My result seem to have more forecast error than of Professor Tatum’s.

6/150 error rate for Tatum’s (see Lecture Notes 6, Tatum, pg. 10).

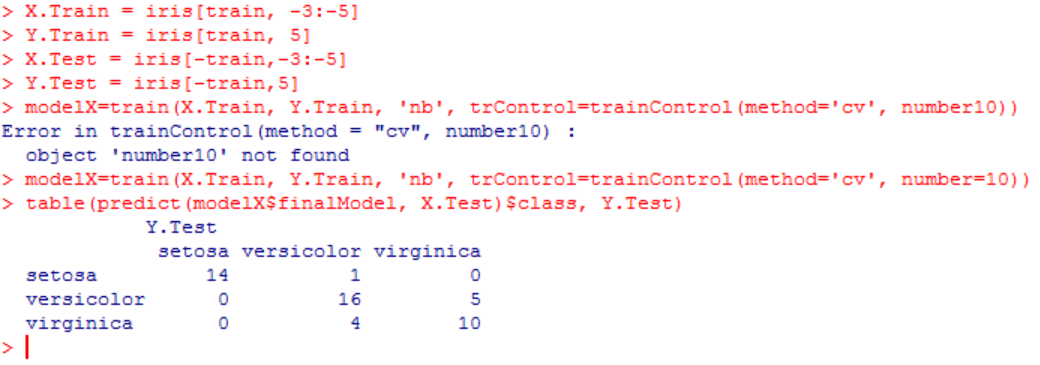
So, I decided to rerun it, and got the same result.

Using sepal length and sepal width as my two

Variables, I got an error rate of 33/150

A higher error rate using 2 variables instead of 4

Is likely a result of using less information to classify species. When 4 variables are taken into account, classification of species yields a better forecast, thus lower forecast error rate.



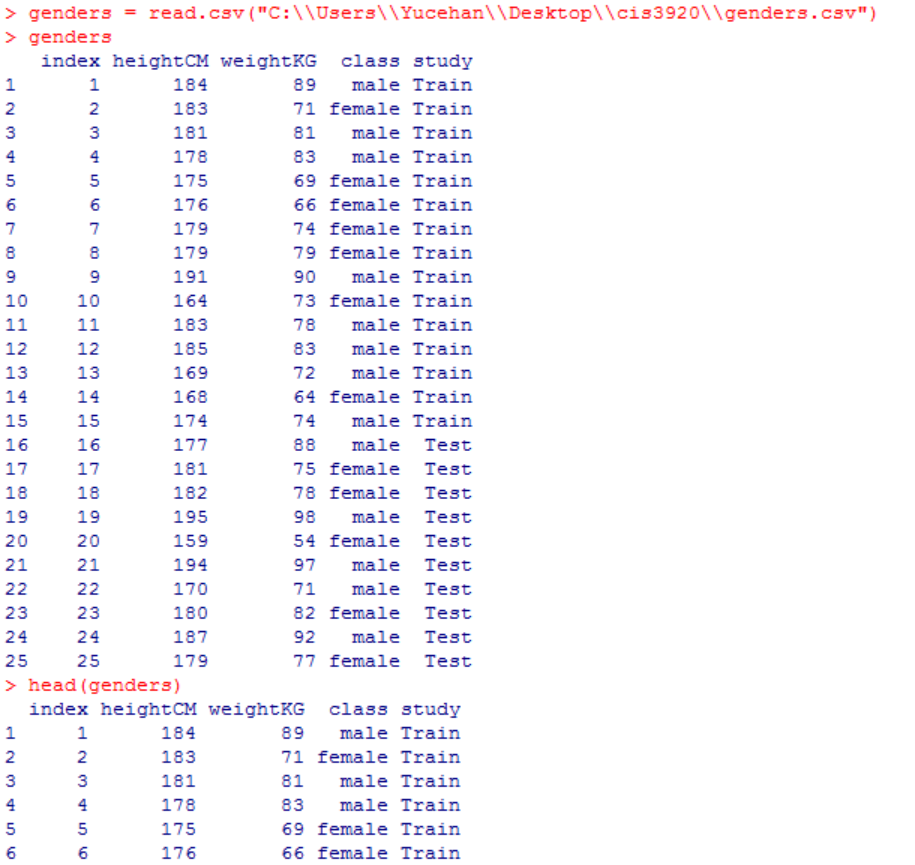
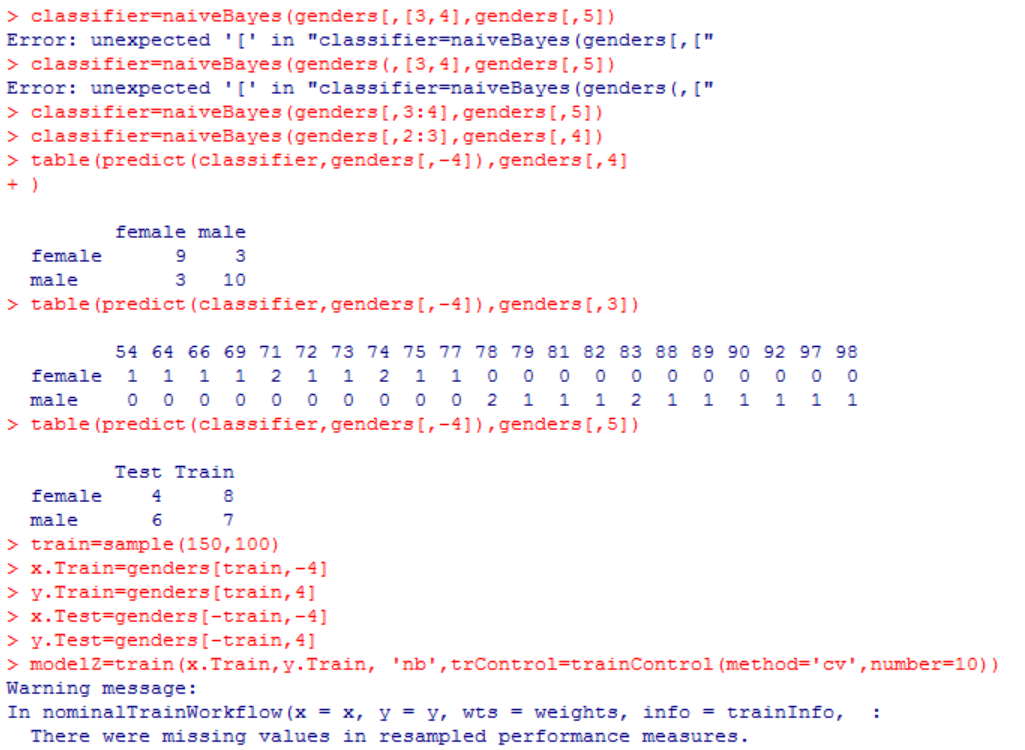
6.2) # assigned X.Train, Y.Train, X.Test, Y.Test to corresponding columns and rows. That is, using only sepal length and width as my two variables.

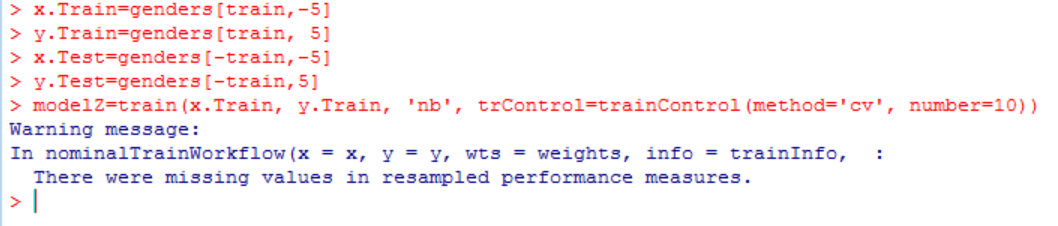
I’ve chosen different variable names to distinguish lecture notes from my own work (X,Y, and T’s are made capital for a slight distinction).

Lawrence using 4 variables, classification error rate (Tatum, Lecture Notes 6, pg. 11) = 2/50=.04

Yucehan using 2 variables (Sepal.Length and Sepal.Width), classification error rate = 10/50 = .2

Using 4 variables yields better classification outcome. 20% error rate is quite high as opposed to 4% which is much better for classification.

6.3)

read my genders.csv file into the R consisting of 25 rows and 5 columns (15 Train, 10 Test).

# could not model it.

6.4) D=event that there is a disease present

P=event that the test is positive

P(D)=1/500; P(P^c|D)=0.02; P(P|D^c)=0.04

=> P(P|D) = 1 – P(P^c|D) = 1 – 0.02 = .98

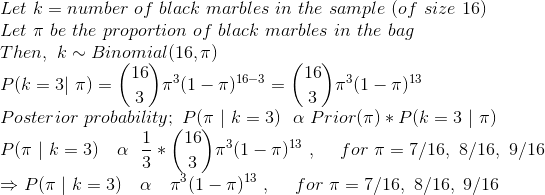
Given a positive test, the probability the disease is present

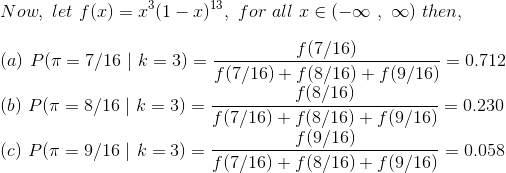
= P(D|P)

= P(D n P)/P(P)=(P(P|D).P(D))/[(P(P|D).P(D))+(P(P|D^c).P(D^c))] (Bayes Theorem)

= (.98 x 1/500)/((.98 x 1/500)+(0.04 x (1-1/500)))

~ 0.0468

6.6)

6.7)  π122(Virginica) means that the probability of a randomly chosen observation that derives from the 122th class’ species will be Virginica.

6.5)

*P(π=0.5|k=3)*

= (P(k=3|π=0.5) x P(π=.5)) / [(P(k=3|π=0.4) x P(π=.4)) + (P(k=3|π=0.5) x P(π=.5)) + (P(k=3|π=0.6) x P(π=.6))]

= 5C3 0.53 0.52 (1/2) / (5C3 0.43 0.62) x ( ¼ ) + 5C3 0.53 0.52 ( ½ ) + 5C3 0.63 0.42 (¼)

= (5C3 ) (0.125\*0.25\*0.5)/[(0.064\*0.36\*0.25) +(0.125\*0.25\*0.5)+(0.216\*0.16\*0.25)]

= 0.015625/ 0.030025 = 52.04%

*P(π=0.6|k=3)*

= (P(k=3|π=0.6) x P(π=.6)) / [(P(k=3|π=0.4) x P(π=.4)) +( P(k=3|π=0.5) x P(π=.5)) + (P(k=3|π=0.6) x P(π=.6))]

= [(5C3) 0.63 0.42 (¼)] / 0.030025

= 0.00864 / 0.030025 = 28.78%

P(π=0.4|k=3) + P(π=0.5|k=3) + P(π=0.6|k=3) = 0.1918 + 0.52039 + 0.2877 = 1 (100%)