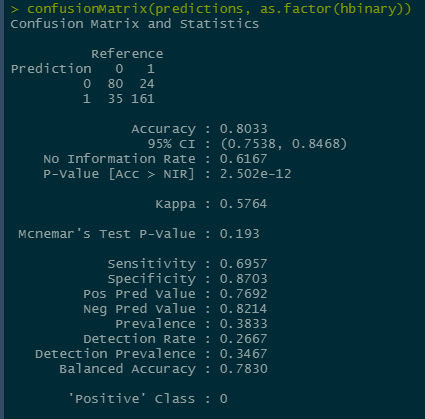
**HOMEWORK 6**

1. On the machine learning portion of HW4, you obtained prediction models with two different approaches, one was provided to you and another you researched. Using your results from HW4 now obtain:

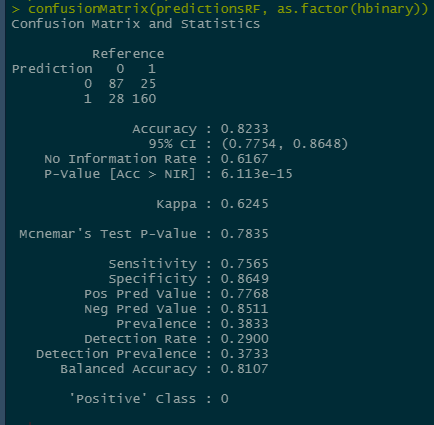
**Please see attached R code for my work for completing this problem.**

1. Confusion Matrix (10 pts)

For the naiveBayes() model:



For the randomForest() model:



1. Performance measures for both methods ( i.e. precision, sensitivity, etc) (20 pts)

See above for accuracy, sensitivity, and specificity.

F-measure = 2\*Specificity\*Neg Pred Value / (Specificty + Neg Pred Value)

F-measure (Naïve Bayes) = 2\*0.8703\*0.8214 / (0.8703 + 0.8214) = 0.8451

F-measure (Random Forest) = 2\*0.8649\*0.8511 / (0.8649 + 0.8511) = 0.8609

c) Analyze your results from a) and b) (20 pts)

The random forest model shows greater accuracy, indicating that it predicted the true values better than the naïve Bayes model. However, the naïve Bayes model showed greater specificity, indicating that out of all of the values that were actually positive, this model predicted more of them correctly than the random forest model. The random forest model also showed great precision, also called the negative prediction value here. This metric indicates that the random forest model has greater correct prediction of positive cases out of all of the cases that were predicted positive. The F-measure is a metric that helps compare two models and is especially useful when the two models have low precision and high recall or vice versa. While this is not the case, I still calculate the F-measure to provide robust metrics. This metric also demonstrates the random forest superiority with random forest having a higher value than naïve Bayes. I think these metrics strongly support that random forest is a better predictive model for this data set. The only metric that was lower for random forest was specificity, however the difference in specificity of the two models was minimal compare to the differences of the other metrics between the two models.

2. Review Ramanathan’s paper (see webliography "Information-theoretic gene-gene and gene-environment interaction analysis of quantitative traits") (50 pts)

- Underline main points of the paper.  
- Keep your work structured.  
- While focusing on big picture keep in mind our class is on statistical processes.  
- Limit to ~1.5/2 pages

In their paper, Chanda et al. (2009) detail their research in generating new interaction analysis methods and an algorithm to associate gene-gene interactions and gene-environment interactions with quantitative traits. Previous research has only analyzed this information in relation to discrete and binary phenotypes, so this would be a significant step forward in the application of information theoretic methods.

The group first identifies that interaction information can represented by the Shannon entropies of the interactions or variables (later A and B) and quantitative trait or phenotype (later P) being measured by combining these entropies for each variable and the phenotype as well as the joined entropy of the phenotype and discrete variables. These entropies can then be used in two different information frameworks: K-way interaction information (KWII) and phenotype-associated information (PAI). KWII can be represented as follows:

KWII(A,B,P) = -H(ABP) + H(AB) + H(AP) + H(BP) + - H(A) – H(B) – H(P)

PAI relies on the total correlation information (TCI), which demonstrates the relationships among the variables where a TCI of 0 means no information can be deduced about the other variables when one variable is known, and the maximum TCI means the known variable is redundant with the other variables. TCI can be represented as follows:

TCI(A,B,P) = H(A) + H(B) + H(P) – H(ABP)

PAI is then represented by the TCI of the phenotype and the overall variable dependency with the variable interdependencies removed. Using both of these methods is advantageous because they complement each other. For example, KWII is not appropriate for something like hill climbing algorithms, but PAI makes up for this because it can be used in such a situation.

Next, the group developed an algorithm, CHORUS, for calculating these information frameworks. First, the PAI is calculated for each variable with the phenotype, and the highest values are stored. This calculation is conducted several times, each time adding a new variable. So the second iteration calculates PAI for two variables and the phenotype, again storing the highest values. For each combination that is stored, the KWII is calculated to indicate which combinations are interacting combinations. In the end, the program returns the interacting combinations of variables and the phenotype.

The group found that CHORUS complements dimensionality reduction methods, because it is less computationally expensive with strong sensitivity, but the algorithm is not as sensitive to specific interactions. In addition, there seems to be a happy medium between the number of high values that are reserved from the PAI calculations and the computational cost of the program to obtain the best results. This makes sense because the more values that are stored, the more active memory that is needed to store the values as well as perform the iterative calculations for those values. However, the algorithm has proven advantageous in dealing with large data sets, shown by the data set used in the study that was approximately 10,000 data points. One weakness of the algorithm that still needed to be addressed at the time of this paper being published is that it is unable to manage continuous covariates. I would expect that further development of this feature would significantly increase the computational cost because there are many more combinations to be considered.

As you can see, Chanda et al. have made great strides in predicting gene-gene and gene-environment interactions as they relate to quantitative traits. The applications of these predictions could see vast improvements in the research of diseases expressing quantitative traits, to lead scientists further down the road to cures for some of the more difficult diseases.