**Homework 4**

1. **This portion of the homework is in the lab style, you will follow the steps to produce predictions with one method and you will research and produce predictions on any other method of your choice (55 points)**

**Part A ( 35 points)**

Machine learning exercise: Follow these steps to make predictions on the following data set :

>install.packages("e1071")

>library(e1071)

>credit\_training <- read.csv("credit\_training.csv")

> credit\_test<-read.csv("credit\_test.csv")

>history\_test<- read.csv("history\_test.csv")

> history\_training<- read.csv("history\_training.csv")

# NOTE that myuniquenumber is an interger between 300 & 600

# you must pick the value and set that variable

> training<-credit\_training[1:myuniquenumber,]

> history<-history\_training[1:myuniquenumber,]

> clsf<-naiveBayes(training,history,0)

> predictions<-predict(clsf,credit\_test)

> table(predictions)

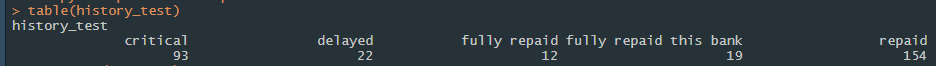
#copy and paste the output

> table(history\_test)

#copy and paste output

Outputs:





**Part B ( 20 points)**

Repeat the process in part B, but use a different algorithm, you may use any of the algorithms presented in the Machine Learning discussion…



1. **Provide 1-1.5 page review for the paper “Genome-wide genetic marker discovery and genotyping using next-generation sequencing” available under this week’s course content (35pts)**

Some guidelines:  
- Underline main points of the paper.  
- Keep your work structured.  
- While focusing on big picture keep in mind our class is on statistical processes.

In their paper, Davey et al. (2011) discuss new advances to next-generation sequencing (NGS) that enhance the discovery of genetic markers. These enhancements are evident through both the development of new methods and changes in the experimental design.

The NGS methods that have developed to improve genetic marker discovery include reduced-representation sequencing, RAD-seq, and low coverage sequencing. Reduced-representation sequencing has been enhanced by the use of RRLs and CRoPS to sample and sequence a small set of genome-wide regions. These methods eliminate the need to sequence the entire genome, saving both time and computing power. In fact, CRoPS has proven to be an improvement on RRLs because the method is able to identify polymorphisms in individuals that are part of a population pool of DNA samples. In addition, the method has been used to evaluate FST statistics in several butterfly populations to determine the differentiation within the populations.

In RAD-seq, short regions near restriction sites for the chosen restriction endonuclease are sequenced, instead of the entire genome. In addition, this method allows for several steps to be carried out using the pooled library, decreasing time and cost. The advantage of this method is having a high density of genome-wide markers without the bulk of data produced by sequencing larger regions.

With low coverage sequencing, the target markers are sequenced at low coverage per individual. Known marker positions and parental genotypes are then used to infer the most likely location for other target markers. In one study, a hidden Markov model was used to assign haplotype based on the probabilities of parental genotypes.

Experimental design has also developed greatly to improve upon marker discovery. As expected, the availability and quality of the reference genome affects the ability to map markers. With a higher quality reference, it’s easier to assign missing genotypes to markers because more of the markers are known. In addition, the expected degree of polymorphisms affects the experimental approach. Fewer polymorphisms means you will want more markers and higher coverage, so it is important to accurately predict the level of polymorphisms so that you can appropriately design the experiment and choose the proper method. Finally, the data analysis end of the experiment has some important factors to consider that are being improved upon. The most notable is the degree of variance that comes from the data. NGS sequencing allows us to process more individuals with more reads, but this increases the variance in the data. However, there are now some available packages to account for this problem through statistical analysis of the data. There are still more improvements that can be made to solve this issue, as the packages are limited to specific methods.

Clearly, many developments have been made to significantly enhance the discovery of genetic markers. These developments are quite often reliant on statistical processes to improve this area, and it seems that the is much room for improvement in the data analysis of experiments to solve some lingering issues that have arisen.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews, 12*, 499-510.

1. **Use <** **HW3 review (Markov Chain and HMM walk through)> file available under course content for this week to write and submit R-script which will (10pts):**
   1. Define HMM model for Q4 in Homework 3
   2. Parse the Homework 3 Q4 sequence to show sequence of hidden states using Viterbi algorithm

See attached R script. The script for question 3 is at the bottom, as I worked both questions 1 and 3 out in R studio in the same script.