CS 640
Data Analysis using R
Relation of SNPs to disease

Single Nucleotide Polymorphisms (SNPs) may be quantified as predictor variables in order to look for associations with a given phenotype, such as a disease. Investigate the dataset disease.arff (on Canvas) containing genomic data on 2000 individuals, some of whom have the disease and some of whom do not. Each row contains data on one individual. There are 10 predictor variables: these are SNPs encoded as number of mutated alleles for the gene. The class variable indicates for each person whether they have disease (1 if they have it, zero if not). The data is in Attribute-Relation File Format (arff), a format used by machine learning packages.

Continue with the dataset that you modeled using Python and scikit-learn. For this assignment explore and analyze the data using R. Fit several models to the data as described below.

Write a report in which you compare your explorations and results and summarize your conclusions about the relationship of the SNPs to the disease. Your report should include plots that convey much of the information. The text you write for the report should be brief. You can paste relevant outputs into your report (ex: summary information) and refer to them to support your conclusions. Use an appendix for lengthy outputs. You may create a jupyter notebook for this assignment. Alternatively you may write a python script and submit your report in Word, pdf, or html format.

Please complete the following tasks:

- Preprocess the data as you see fit (ex: R or Python).
- Use R to explore the data. Produce summary statistics and plots to visualize the data.
- Fit regression models in R to model disease using the predictor variables. Your models should include main effects only, interactions only and the full model. Produce a final model containing all the predictors you find that are significant.
- In your report evaluate each of your models. State how the model is affected by a 1-unit increase in each predictor variable when the other predictor variables are held constant. Produce a plot for each model.