Assignment 4 Writeup

Group: 5, Date: 11/06/24

Hannah Luft

Code: KNN.R

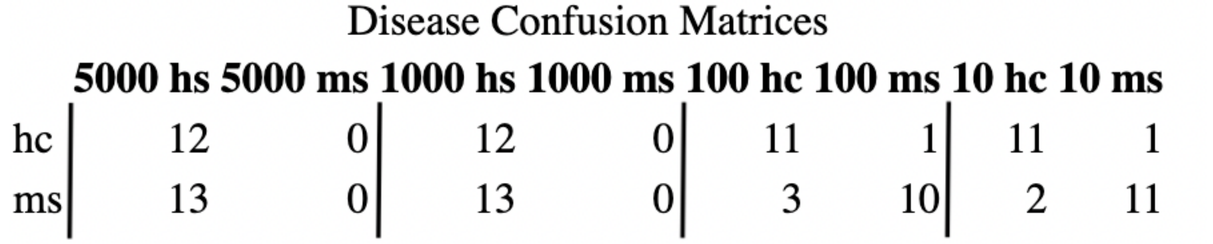
Supervised Analysis: K Nearest Neighbors

For this assignment, the K Nearest Neightbors (KNN) algorithm was run on the top 5000 most variable genes from our dataset. It was used to predict the disease groups (healthy control(hc) vs multiple sclerosis(ms)), and the clusters from Consensus Cluster Plus, at both k=2 and k=3 (two and three cluster groups).

The algorithm was trained using a random set of 60% of the samples. The training sample set was kept consistent for the purposes of direct comparison of performance (they were trained with the same expression data from the same samples each time) between scenarios *and* machine learning algorithms each member of the group used.

Results and Confusion Matrices

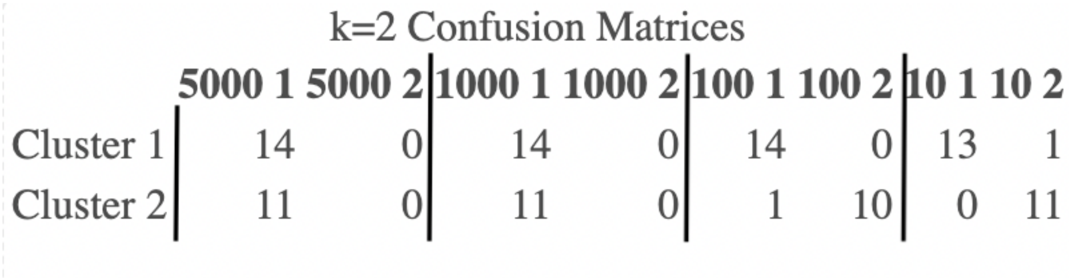
KNN was trained and tested with 5000 genes, 1000 genes, 100 genes, and 10 genes, for each of the three above stated group predictions (disease, k=2, and k=3). Interestingly, the KNN algorithm performed better with fewer genes. We suspect that this is because our dataset has very few genes with significant differential expression and variance, so adding more gene data is actually just introducing bad training data, wherein most of the expression data is just 0. Thus, the fewer genes, the better the training data, and the better the results. The confusion matrices are shown below, grouped by disease, k=2 clusters, and k=3 clusters; sub-grouped by gene amounts (5000, 1000, 100, and 10).



Disease Confusion Matrices

	5000 hs	5000 ms	1000 hs	1000 ms	100 hc	100 ms	10 hc	10 ms
hc	12	0	12	0	11	1	11	1
ms	13	0	13	0	3	10	2	11

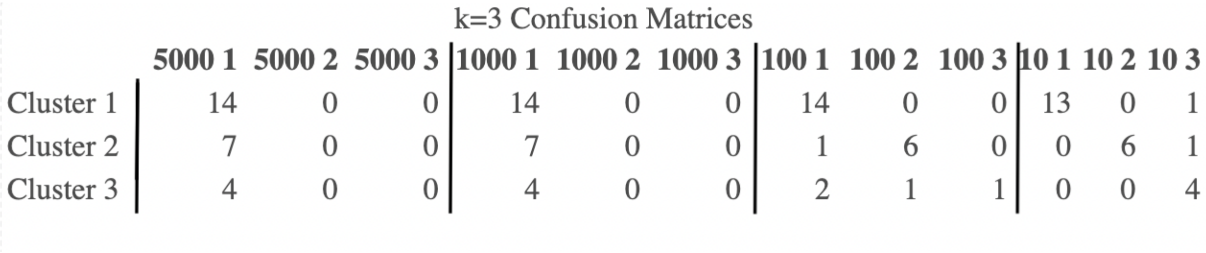
Figure 1: Disease Group Confusion Matrices



k=2 Confusion Matrices

	5000 1	5000 2	1000 1	1000 2	100 1	100 2	10 1	10 2
Cluster 1	14	0	14	0	14	0	13	1
Cluster 2	11	0	11	0	1	10	0	11

Figure 2: Cluster k=2 Group Confusion Matrices



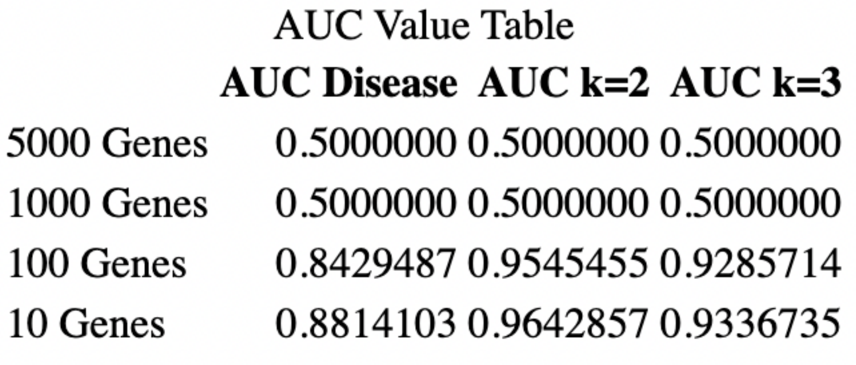
k=3 Confusion Matrices

	5000 1	5000 2	5000 3	1000 1	1000 2	1000 3	100 1	100 2	100 3	10 1	10 2	10 3
Cluster 1	14	0	0	14	0	0	14	0	0	13	0	1
Cluster 2	7	0	0	7	0	0	1	6	0	0	6	1
Cluster 3	4	0	0	4	0	0	2	1	1	0	0	4

Figure 3: Cluster k=3 Group Confusion Matrices

ROC AUC

The area under the curve (AUC) was calculated for each scenerio's receiver operating characteristic (ROC) curve using the pROC library. The AUC results are shown below. Matching the KNN predictions and confusion curves, the KNN models did better with smaller numbers of genes. For 1000 and 5000 genes, the AUC values were 0.5, indicating that it is no better at predicting the groups than chance. This supports the bad input data argument. Additionally, the KNN model was best at predicting the cluster groups for k=2, more so than predicting the actual disease groups.



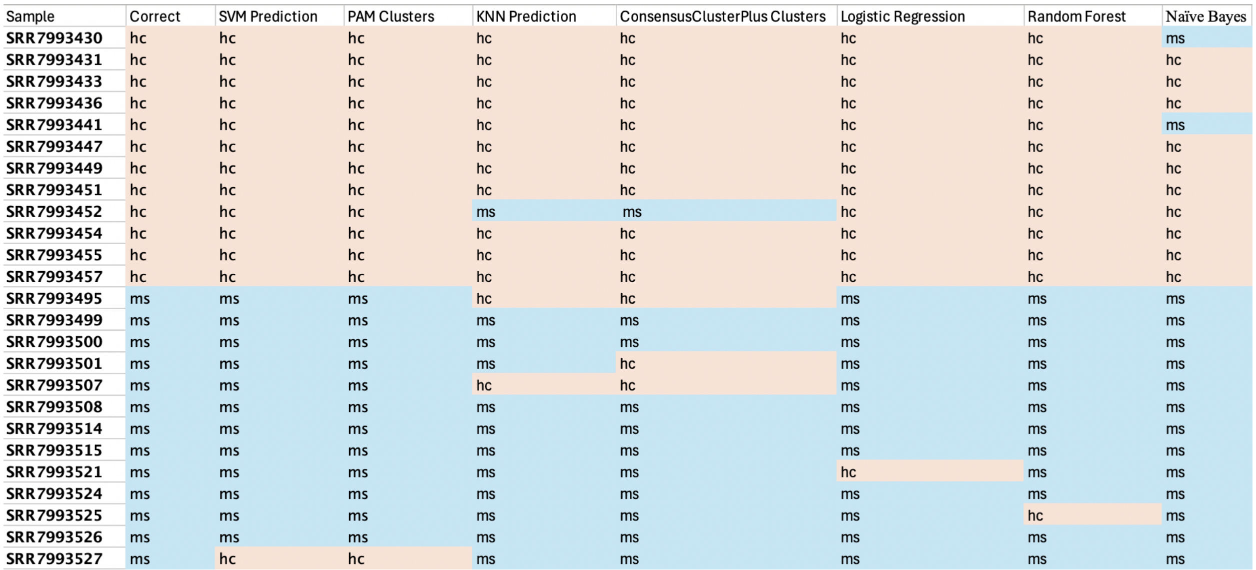
AUC Value Table

	AUC Disease	AUC k=2	AUC k=3
5000 Genes	0.5000000	0.5000000	0.5000000
1000 Genes	0.5000000	0.5000000	0.5000000
100 Genes	0.8429487	0.9545455	0.9285714
10 Genes	0.8814103	0.9642857	0.9336735

Figure 4: AUC Table

Group Comparison of Models

The results of each group member's unsupervised clusters and supervised predictions for the test set of samples were tabulated. My unsupervised clustering method was Consensus Cluster Plus, and as mentioned my supervised analysis algorithm was K Nearest Neighbors. The KNN results listed are my best results, from running KNN with 10 genes on the disease groups, and the Consensus Cluster Plus results are from the k=2 clusters also at 10 genes. While neither algorithm was able to perfectly predict disease groups, they were fairly consistent with each other, with only one mismatching sample.



Sample	Correct	SVM Prediction	PAM Clusters	KNN Prediction	ConsensusClusterPlus Clusters	Logistic Regression	Random Forest	Naïve Bayes
SRR7993430	hc	hc	hc	hc	hc	hc	hc	ms
SRR7993431	hc	hc	hc	hc	hc	hc	hc	hc
SRR7993433	hc	hc	hc	hc	hc	hc	hc	hc
SRR7993436	hc	hc	hc	hc	hc	hc	hc	hc
SRR7993441	hc	hc	hc	hc	hc	hc	hc	ms
SRR7993447	hc	hc	hc	hc	hc	hc	hc	hc
SRR7993449	hc	hc	hc	hc	hc	hc	hc	hc
SRR7993451	hc	hc	hc	hc	hc	hc	hc	hc
SRR7993452	hc	hc	hc	ms	ms	hc	hc	hc
SRR7993454	hc	hc	hc	hc	hc	hc	hc	hc
SRR7993455	hc	hc	hc	hc	hc	hc	hc	hc
SRR7993457	hc	hc	hc	hc	hc	hc	hc	hc
SRR7993495	ms	ms	ms	hc	hc	ms	ms	ms
SRR7993499	ms	ms	ms	ms	ms	ms	ms	ms
SRR7993500	ms	ms	ms	ms	ms	ms	ms	ms
SRR7993501	ms	ms	ms	ms	hc	ms	ms	ms
SRR7993507	ms	ms	ms	hc	hc	ms	ms	ms
SRR7993508	ms	ms	ms	ms	ms	ms	ms	ms
SRR7993514	ms	ms	ms	ms	ms	ms	ms	ms
SRR7993515	ms	ms	ms	ms	ms	ms	ms	ms
SRR7993521	ms	ms	ms	ms	ms	hc	ms	ms
SRR7993524	ms	ms	ms	ms	ms	ms	ms	ms
SRR7993525	ms	ms	ms	ms	ms	ms	hc	ms
SRR7993526	ms	ms	ms	ms	ms	ms	ms	ms
SRR7993527	ms	hc	hc	ms	ms	ms	ms	ms

Figure 5: Group Results Comparison

Heatmap and Dendrogram

A heatmap with dendrograms was created from the results using the Complex Heatmap library. It contains annotation bars of the true disease groups, the predicted disease groups, predicted cluster groups for k=2, and predicted cluster groups for k=3. The top 1000 genes by variance are plotted because the heatmap would not render with a higher number; additionally, the most varied and thus most relavent genes are all contained within the top 100 genes, so adding the next 4000 would not add particularly important information, only more blue (0) values.

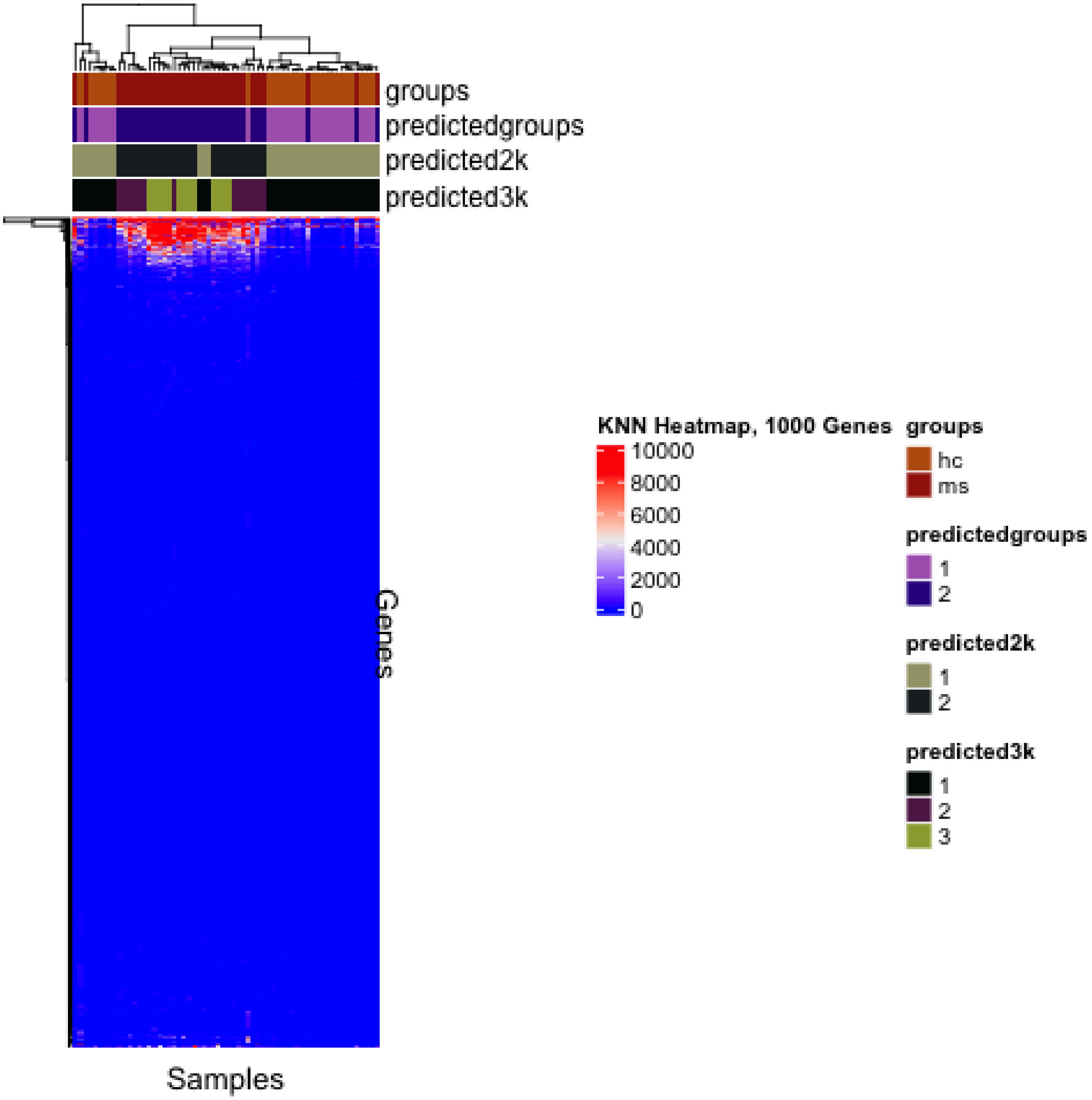


Figure 6: KNN Heatmap