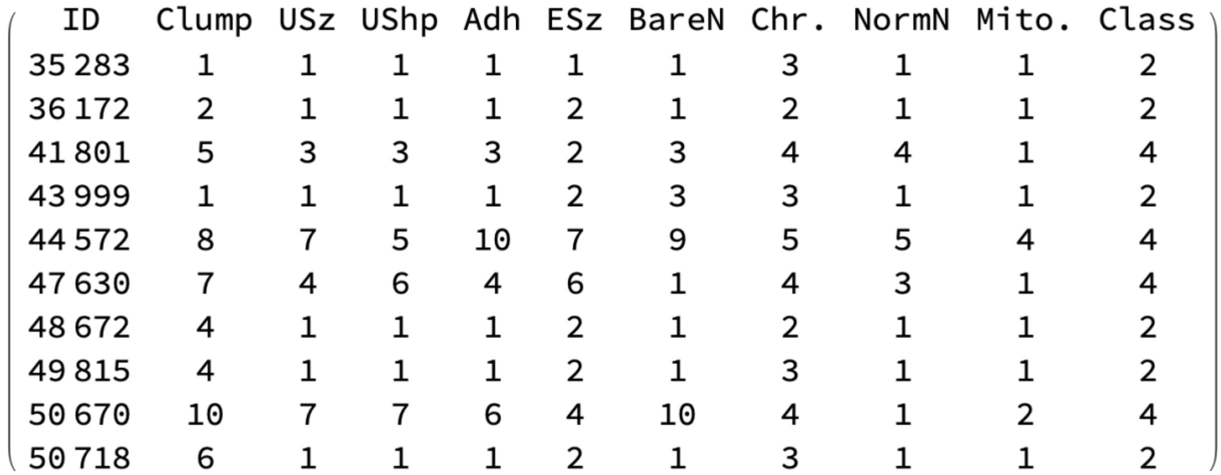
CSE514, Fall 2021, HW 7 Name: Anh Le Student ID: 488493

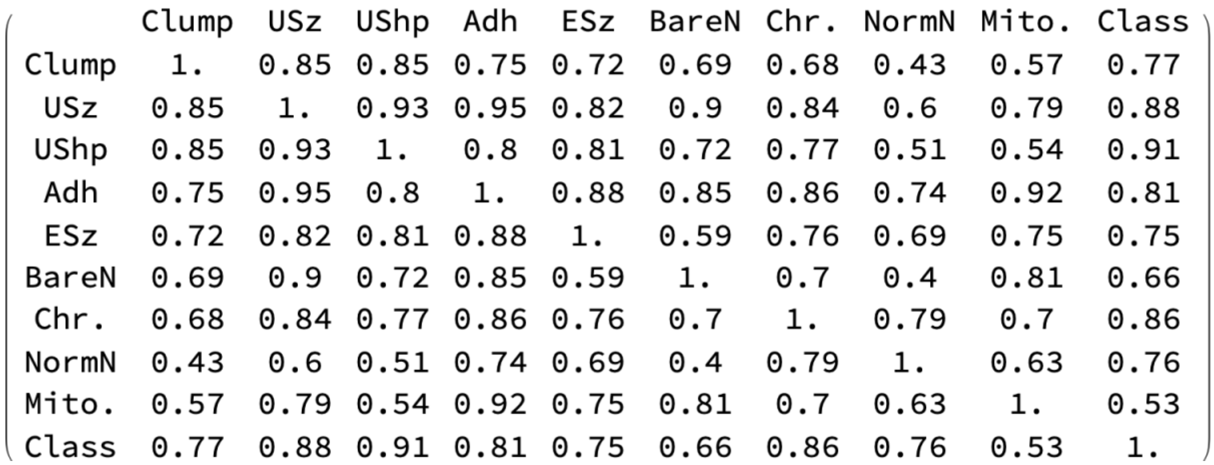
Note: This homework is worth a total of 15 points

Here are samples 11-20 of the breast cancer data from your Programming Assignment 2:
(*ID numbers had the first three digits of 10 dropped to save space)



ID	Clump	USz	UShp	Adh	ESz	BareN	Chr.	NormN	Mito.	Class
35 283	1	1	1	1	1	1	3	1	1	2
36 172	2	1	1	1	2	1	2	1	1	2
41 801	5	3	3	3	2	3	4	4	1	4
43 999	1	1	1	1	2	3	3	1	1	2
44 572	8	7	5	10	7	9	5	5	4	4
47 630	7	4	6	4	6	1	4	3	1	4
48 672	4	1	1	1	2	1	2	1	1	2
49 815	4	1	1	1	2	1	3	1	1	2
50 670	10	7	7	6	4	10	4	1	2	4
50 718	6	1	1	1	2	1	3	1	1	2

Here's matrix of the correlation between all the features of this subset, including the response value of Class:



	Clump	USz	UShp	Adh	ESz	BareN	Chr.	NormN	Mito.	Class
Clump	1.	0.85	0.85	0.75	0.72	0.69	0.68	0.43	0.57	0.77
USz	0.85	1.	0.93	0.95	0.82	0.9	0.84	0.6	0.79	0.88
UShp	0.85	0.93	1.	0.8	0.81	0.72	0.77	0.51	0.54	0.91
Adh	0.75	0.95	0.8	1.	0.88	0.85	0.86	0.74	0.92	0.81
ESz	0.72	0.82	0.81	0.88	1.	0.59	0.76	0.69	0.75	0.75
BareN	0.69	0.9	0.72	0.85	0.59	1.	0.7	0.4	0.81	0.66
Chr.	0.68	0.84	0.77	0.86	0.76	0.7	1.	0.79	0.7	0.86
NormN	0.43	0.6	0.51	0.74	0.69	0.4	0.79	1.	0.63	0.76
Mito.	0.57	0.79	0.54	0.92	0.75	0.81	0.7	0.63	1.	0.53
Class	0.77	0.88	0.91	0.81	0.75	0.66	0.86	0.76	0.53	1.

Q1 (3pts): For the end goal of training a classifier on these 10 samples, give **two** reasons why reducing the number of features would be a good idea.

- The more feature variables, the higher the computation cost of modeling.
- The number of samples (10) is approximately the same as the number of features (9) so if we keep all features, more samples are required to fit an accurate model.

Q2 (6pts): Filter methods

Q2a (3pts): What is an **unsupervised** filter method that you could apply using the given correlation matrix? Use this method to reduce the dimensions of this dataset by 1.

- When a pair of features are predictive of each other, get rid of one. In this case, correlation point between USz and Adh is high (0.95), meaning that the samples are relatively same. We only see that sample 44572 differs a lot (7 vs 10) between USz and Adh so we can consider dropping one.

Q2b (3pts): What is a **supervised** filter method that you could apply using the given correlation matrix? Use this method to reduce the dimensions of this dataset by 1.

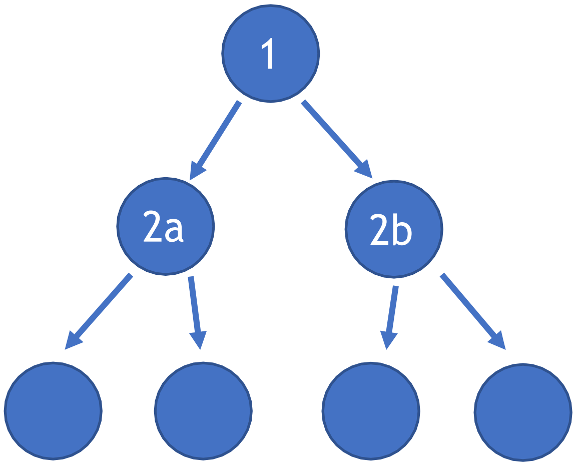
- The feature that has the lowest correlation point to the class output compared to other features can be dropped . So in this case, I would drop Mito feature (0.53).

Q3 (4pts): Embedded methods.

Imagine I trained a random forest of 10 trees on a similar subset of the breast cancer dataset. Each tree is fit to 3 random features (highlighted green) and a bootstrapped dataset.
Here's how the 10 trees picked features to split on:

	Clump	USz	UShp	Adh	ESz	BareN	Chr.	NormN	Mito.
1				1	2b	2a			
2		1	2b			2a			
3	2b				1	2a			
4	2a			1				2b	
5				2a		1			2b
6	2a, 2b		🟩						1
7	2b		2a			1			
8			🟩	2a, 2b	1				
9		1	2a				2b		
10				1	2a, 2b		🟩		

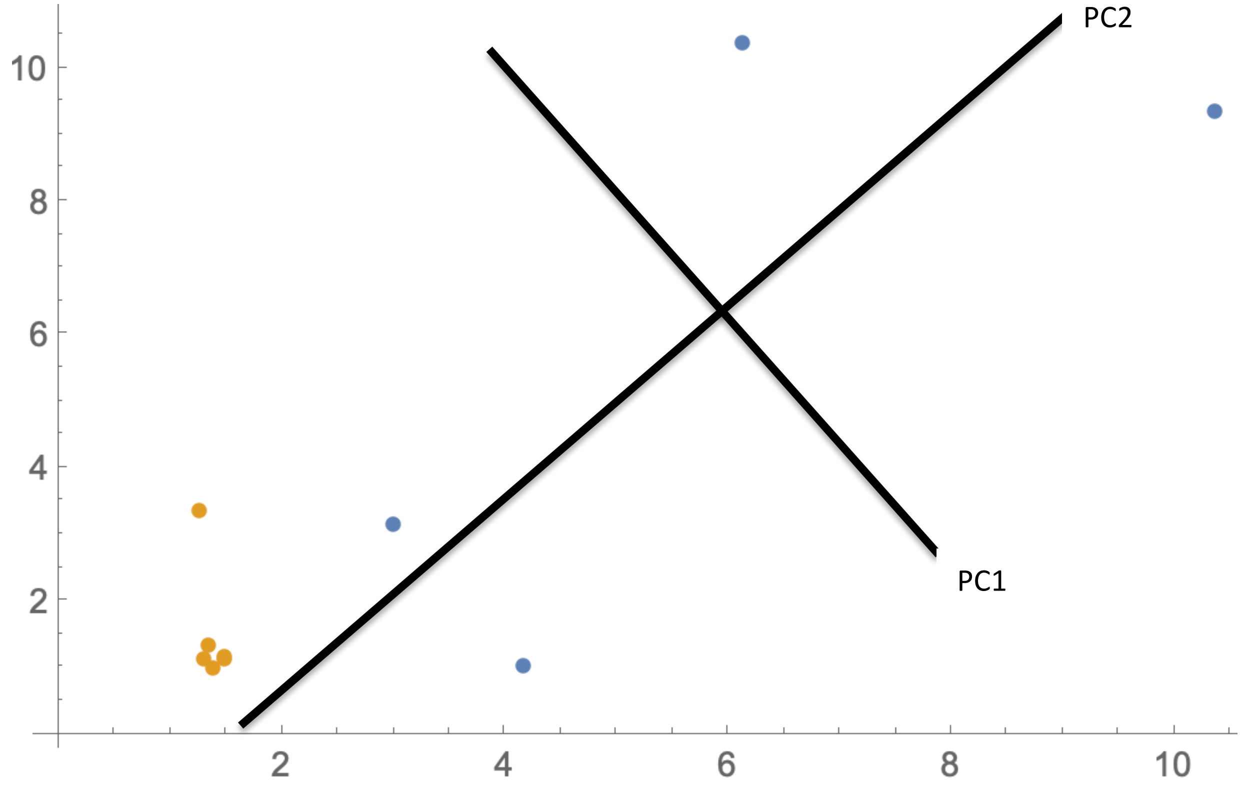
1: indicates the first split at the root
2a: indicates the split at the right child of the root
2b: indicates the split at the left child of the root



Which 3 features would you choose to drop based on this random forest?
Explain why.

- I would drop Chr., NormN, and USz because these features do not indicate much split at either root or children of the trees. Since the features chosen for splitting are picked to maximize decrease in Gini Impurty or Varian, Chr., NormN, and USz would probably not maximize decrease in GI, hence consider dropping them.

Q4 (2pts): I've plotted the dataset using only the Adh and BareN features:



Class = 4 is in blue and Class = 2 is in yellow. Jitter was used so that all the points could be seen. Draw on the plot your best guess for PC1 and PC2. Make sure to label which is which.

EC (2pts): Use any programming language and/or libraries to calculate the first two principal components of the dataset using all 9 features and 10 samples.
Submit a plot of the data using PC1 as the x-axis and PC2 as the y-axis.
Use different colors or shapes for samples from different classes, like was done for **Q4.**