Answers for Q1

(a) 292
(b) Data mining
(c)

	min	max	mean	s.d
Mcg	0	0.89	0.4664	0.1828
Alm1	0.03	1.0	0.4679	0.2095

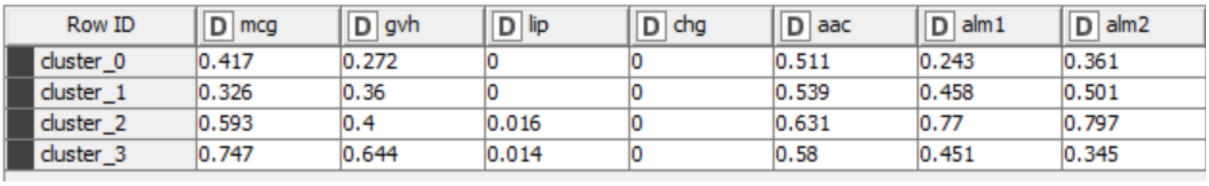
(d)

	Mean		s.d	
	mcg	gvh	mcg	gvh
Cp	0.364	0.41	0.124	0.09
Im	0.478	0.497	0.195	0.088
Om	0.672	0.71	0.069	0.12
Pp	0.652	0.7	0.09	0.129

(e) Most proteins located at cp (48.97%) and im (26.37%)
(f) Mcg for cp appear to be smaller compared to pp. mcg for cp has a symmetric distribution while mcg for pp is negatively skewed.
(g) ‘gvh’ alone not sufficient to classify because boxplot overlaps
(h) Yes, scatterplot show two well-separated clusters for cp and im.
(i) Alm1 and alm2.
(j) r=0.79611, strong positive correlation.

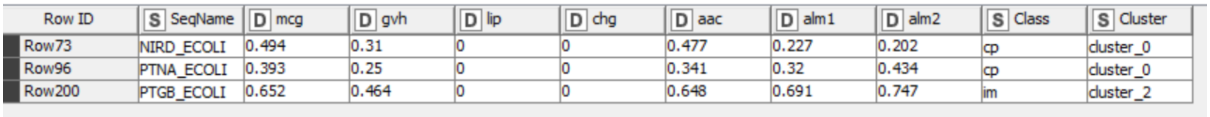
Answers for Q2

(a) K is 4 because we are clustering into 4 classes.
(b) Normalization prevents any attribute with large numerical values from dominating the distance calculations.
(c)



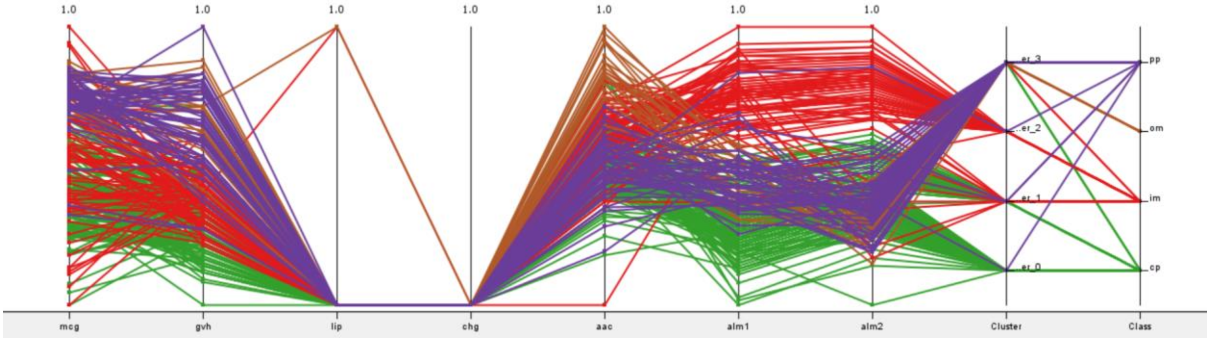
Row ID	D mcg	D gvh	D lip	D chg	D aac	D alm1	D alm2
cluster_0	0.417	0.272	0	0	0.511	0.243	0.361
cluster_1	0.326	0.36	0	0	0.539	0.458	0.501
cluster_2	0.593	0.4	0.016	0	0.631	0.77	0.797
cluster_3	0.747	0.644	0.014	0	0.58	0.451	0.345

(d) Cluster 0 with 99 records.
(e)



Row ID	S SeqName	D mcg	D gvh	D lip	D chg	D aac	D alm1	D alm2	S Class	S Cluster
Row73	NIRD_ECOLI	0.494	0.31	0	0	0.477	0.227	0.202	cp	cluster_0
Row96	PTNA_ECOLI	0.393	0.25	0	0	0.341	0.32	0.434	cp	cluster_0
Row200	PTGB_ECOLI	0.652	0.464	0	0	0.648	0.691	0.747	im	cluster_2

(f)



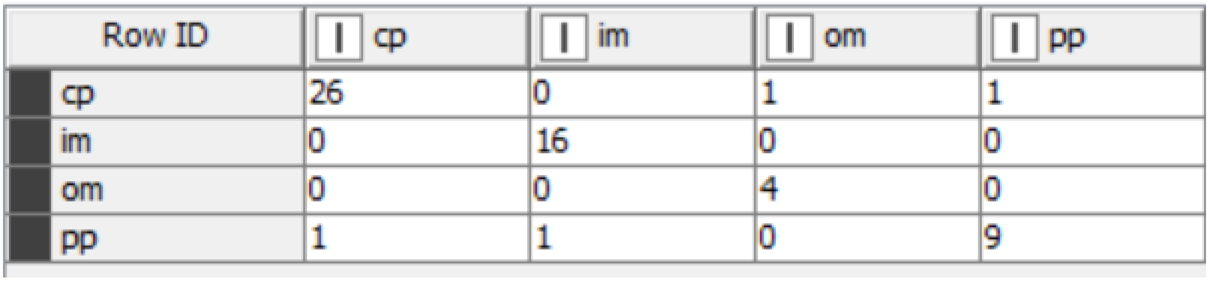
Based on parallel coordinates plot, cp is not well separated. Half of cp went to cluster 0 and the other half went to cluster 1. On the other hand, im seems to be well separated as majority went to cluster 2. cp cluster is characterised by lower than average alm1. im cluster is characterised by higher than average alm1 and alm2.

Answers for Q3

(a) Class

(b) Target is categorical

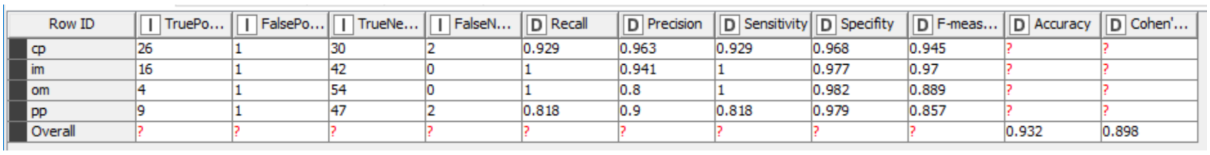
(c)



Row ID	I cp	I im	I om	I pp
cp	26	0	1	1
im	0	16	0	0
om	0	0	4	0
pp	1	1	0	9

Accuracy statistics – 93.2%

(d)



Row ID	I TruePo…	I FalsePo…	I TrueNe…	I FalseN…	D Recall	D Precision	D Sensitivity	D Specifity	D F-meas…	D Accuracy	D Cohen'…
cp	26	1	30	2	0.929	0.963	0.929	0.968	0.945	?	?
im	16	1	42	0	1	0.941	1	0.977	0.97	?	?
om	4	1	54	0	1	0.8	1	0.982	0.889	?	?
pp	9	1	47	2	0.818	0.9	0.818	0.979	0.857	?	?
Overall	?	?	?	?	?	?	?	?	?	0.932	0.898

im and om have no FN.

(e) 2 records from cp are FN – wrongly classified as om and pp.

(f) If alm1>0.6 then class = im with prob 0.951
If alm1<=0.6 and aac>0.64 then class = om with prob 0.875
If alm1<=0.6 and aac<=0.64 and gvh>0.58 then class = pp with prob 0.941
If alm1<=0.6 and aac<=0.64 and gvh<=0.58 then class = cp with prob 0.934

(g) Classified as pp

(h) Prune it

(i) Accuracy improves to 94.9%

(j) Both first split at alm1 with different criteria. Gini splits at alm1 (0.575). Second split for Gini is gvh instead of aac.

Answers for Q4

(a) Yield is response and ph is predictor.

(b) $yield = -950.7 + 334.7826\,ph$

(c) The intercept is -950.7. This is the baseline. Estimated yield is -950.7 units when ph is 0. This is not a reasonable value as yield cannot be negative.
As ph increases by 1 unit, estimated yield is expected to increase by 334.7826 units.
R^2 is 0.6138. Moderately reliable for prediction.

(d) 33.47826units

(e) 723.22

(f) R^2 is quite high. Ok for prediction provided within range of ph from 4.6 to 6.0