SAE EST Revision 1920S1

Questions 1 - 4 require the use of KNIME software and Questions 1 – 3 require the data file “Ecoli.csv”. Below is the description of the “Ecoli.csv”.

This Ecoli dataset (Resource: UCI Machine Learning Repository) contains protein localization sites. The meaning of the attributes are given as follows:

Attribute	**Meaning**
SeqName	Accession number for the SWISS-PROT database
mcg	McGeoch's method for signal sequence recognition
gvh	von Heijne's method for signal sequence recognition
lip	von Heijne's Signal Peptidase II consensus sequence score
chg	Presence of charge on N-terminus of predicted lipoproteins
aac	Score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins
alm1	Score of the ALOM membrane spanning region prediction program
alm2	Score of ALOM program after excluding putative cleavable signal regions from the sequence
Class	Localization site (‘cp’ (cytoplasm), ‘im’ (inner membrane), ‘pp’ (perisplasm), ‘om’ (outer membrane))

Question 1

The data in the Excel file “Ecoli.csv” will be used to build a model for the classification of the localization site into ‘cp’, ‘im’, ‘pp’ and ‘om’.

a) How many records are there in the dataset?

b) Is the task described above a data query task or a data mining task?

c) Find the overall minimum, maximum, mean and standard deviation for the following attributes:

(i) ‘mcg’

(ii) ‘alm1’

d) For each localization site, find the mean and standard deviation for each of the two attributes:

Class	Mean mcg	Mean gvh	Standard deviation mcg	Standard deviation gvh
cp				
im				
om				
pp				

e) Construct a pie chart (or bar graph) for *class*. Comment on your results. (Note: You do not have to draw the chart here.)

f) Construct a histogram for ‘mcg’. Comment on your results for the ‘cp’ and ‘pp’ *class*. (Note: You do not have to draw the chart here.)

g) Construct a box-and-whisker plot showing the distribution of ‘gvh’ for each *class*. Based on your plot, comment on whether ‘gvh’ alone can be used to classify into their respective localization site. (Note: You do not have to draw the chart here.)

h) Construct a scatterplot between the ‘mcg’ and the ‘alm2’ using a colour scheme to distinguish between different localization site in the plot. Do you think these 2 attributes are significant in distinguishing between the localization site ‘cp’ and ‘im’? Explain. (Note: You do not have to draw the chart here.)

i) Construct a scatterplot matrix for the attributes ‘mcg’, ‘gvh’, ‘alm1’ and ‘alm2’. Which pair of attributes suggest a linear relationship?

j) Based on part (i) above, investigate further to compute the pairwise correlations for the variables. Comment on your results.

Question 2

Perform K-Means clustering on the Ecoli dataset to uncover the clusters of its localization sites. In the Partitioning node, choose the size of first partition as relative 99% drawn randomly using the random see of ‘1234’. The first partitioned data is to be normalized using min-max (0.0-1.0) and used to create the K-means model. The second partitioned data will be used for the testing.
For centroid initialization, select the option 'First k rows'.

a) What is the value of k to be used for the K-means model? Explain why.

b) Explain why normalization is important.

c) Write down the resulting cluster centroids for each cluster.

d) Which cluster group contains the most number of records? State the number of records for that group.

e) Based on the model created in (a), write down the cluster ID assigned to the second partitioned data used for testing.

f) Based on the parallel coordinates visualization for the cluster result, comment on the separability of ‘cp’ and ‘im’, and state any attributes that characterize the cluster.

Question 3

Build a classification model to predict the localization sites using the decision tree method with ‘Gain ratio’ (no pruning). Reduced error pruning is checked to cut the tree in post-processing by the algorithm in KNIME. In the Partitioning node, choose the size of first partition as relative 80% stratified sampled using the random seed of ‘1234’. Restrict the minimum number of records per node to be 15.

a) Which is the target variable?

b) Explain why a regression model is not selected here.

c) Make use of the confusion matrix generated by the classification workflow to compute the overall accuracy of the decision tree. Compare that with the accuracy statistics generated by KNIME.

d) From the confusion matrix generated by the classification workflow, which is/are the localization site(s) with no False Negatives produced? Interpret your answer in context.

e) From the confusion matrix generated by the classification workflow, how many False Negative for class ‘cp’ are produced? Interpret your answer in context.

f) Based on the decision tree produced above, summarize the tree as a set of classification rules of the form “If… Then…” for each localization site.

g) Use the decision tree above to predict the localization site of a record with the following attribute values: SeqName = AAT_ECOLI, mcg = 0.5, gvh = 0.6, lip = 0.48, chg = 0.5, acc = 0.6, alm1 = 0.5, alm2 = 0.35.

h) Name one way to avoid overfitting when building a decision tree.

i) Build a classification model to predict the localization site using the decision tree method with ‘Gini Index’, keeping all other parameter settings similar as above.

   a. Observe the model and compare its performance in terms of accuracy of classification with the decision tree above.

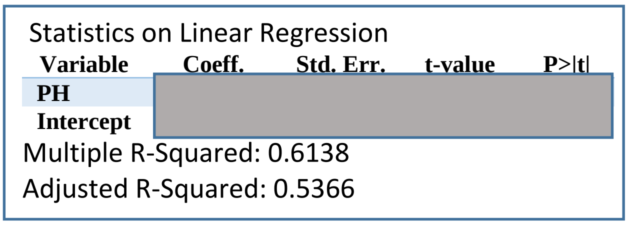
   b. Compare and discuss on the first split and second split of the two decision trees generated.

Question 4

A scientist planted alfalfa on several plots of land, identical except for the soil pH. The data collected and shown below give the yields (in kilograms per acre) for each (refer to Table 1). He obtained the linear regression results with the values masked out (refer to Table 2).

pH	Yield
4.6	479
4.8	831
5.2	739
5.4	840
5.6	809
5.8	1201
6.0	967

Table 1



Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
PH	[redacted]	[redacted]	[redacted]	[redacted]
Intercept	[redacted]	[redacted]	[redacted]	[redacted]

Multiple R-Squared: 0.6138

Adjusted R-Squared: 0.5366

Table 2

(a) Which variable is the response variable and which is the predictor variable?

(b) Build a simple linear regression model using KNIME. Write down the equation of the model.

(c) Interpret the model obtained in (b).

(d) For every increase of 0.1 in soil pH, how much would you expect the yield to change?

(e) Estimate the yield from a plot of land with soil pH 5.0.

(f) Is the equation useful to estimate the yield? Explain in context.