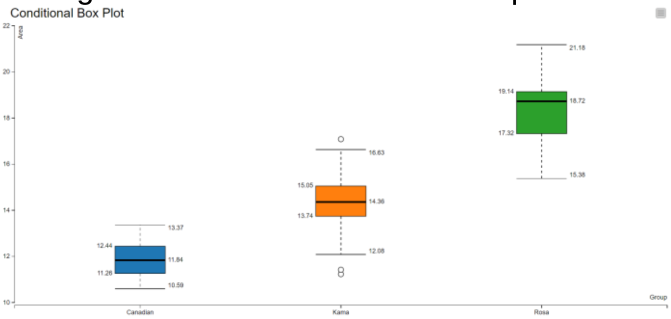
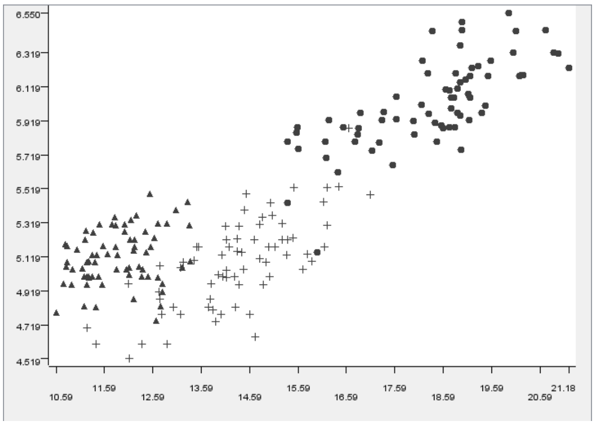
SOLUTIONS

SINGAPORE POLYTECHNIC
2019 / 2020 Semester 1 EST

Module Name: Statistics and Analytics for Engineers **Module Code: MS_SAE**

Course: **Year: 2 FT**

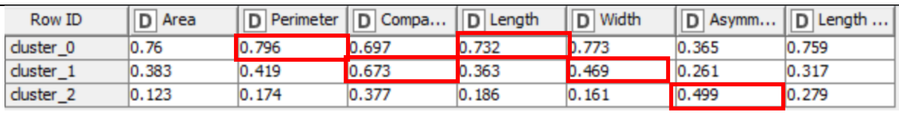
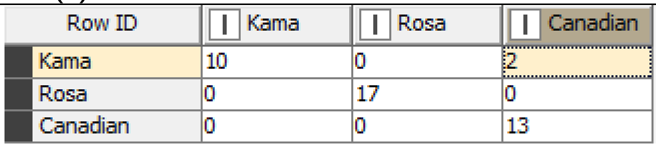
Qn	Solution	Total
1a	210 records	10
1b	Data mining task	
1c	(i) 14.8 (ii) Canadian group mean area: 11.9 Kama group mean area: 14.3 Rosa group mean area: 18.3 (iii) As we see overlaps over the range (Canadian and Rosa seen to have little overlap compared to Canadian and Kama on the same scale from the boxplot), we can conclude that area alone cannot accurately be used to distinguish between different seed species.  (iv) Yes, because the two clusters are distinctly separated as shown in the scatterplot for ‘Canadian’ (triangle) and ‘Rosa’ (circle). ‘Canadian’ generally have small area and length of kernel groove; while ‘Rosa’ generally have large area and length of kernel groove. 	
2a	*k* = 3 because there are three different species of seeds.	

SOLUTIONS

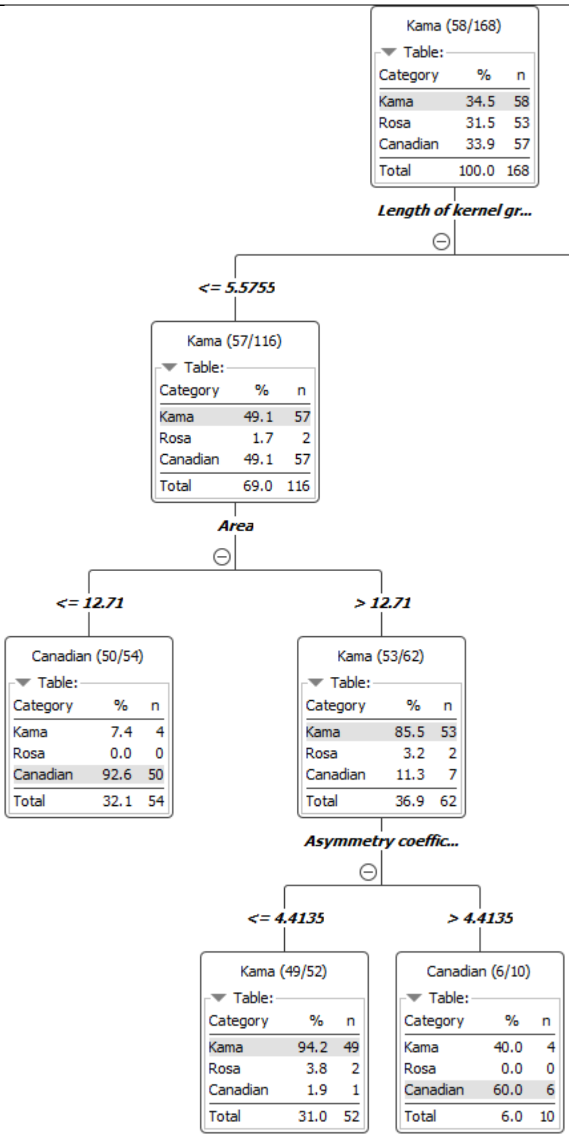
SINGAPORE POLYTECHNIC
2019 / 2020 Semester 1 EST

Module Name: Statistics and Analytics for Engineers **Module Code: MS_SAE**

Course: **Year: 2 FT**

<table>
<tr><td></td><td></td><td rowspan="5">10</td></tr>
<tr><td>2b</td><td>Clustering was done on normalized data so as not to let any attribute dominate the distance calculation due to large numerical values.</td></tr>
<tr><td>2c</td><td>Number of records in cluster_0 = 63
Number of records in cluster_1 = 68
Number of records in cluster_2 = 76</td></tr>
<tr><td>2d</td><td>
<table>
<tr><th>Row ID</th><th>D Area</th><th>D Perimeter</th><th>D Compa...</th><th>D Length</th><th>D Width</th><th>D Asymm...</th><th>D Length ...</th></tr>
<tr><td>cluster_0</td><td>0.76</td><td>0.796</td><td>0.697</td><td>0.732</td><td>0.773</td><td>0.365</td><td>0.759</td></tr>
<tr><td>cluster_1</td><td>0.383</td><td>0.419</td><td>0.673</td><td>0.363</td><td>0.469</td><td>0.261</td><td>0.317</td></tr>
<tr><td>cluster_2</td><td>0.123</td><td>0.174</td><td>0.377</td><td>0.186</td><td>0.161</td><td>0.499</td><td>0.279</td></tr>
</table></td></tr>
<tr><td>2e</td><td>cluster_1
cluster_0
cluster_2</td></tr>
<tr><td>3a</td><td>Group</td><td></td></tr>
<tr><td>3b</td><td>(i) Node: File Reader
(ii) Node: Partitioning
(iii) Node: Decision Tree Learner
(iv) Node: Decision Tree Predictor
(v) Node: Scorer</td><td></td></tr>
<tr><td>3c</td><td>
<table>
<tr><th>Row ID</th><th>I Kama</th><th>I Rosa</th><th>I Canadian</th></tr>
<tr><td>Kama</td><td>10</td><td>0</td><td>2</td></tr>
<tr><td>Rosa</td><td>0</td><td>17</td><td>0</td></tr>
<tr><td>Canadian</td><td>0</td><td>0</td><td>13</td></tr>
</table>
False Negative for ‘Kama’ = 2
It means that 2 records of Kama seed have been misclassified as Canadian seed by the decision tree model.</td><td></td></tr>
</table>

3d



Kama (58/168)

Table:

Category	%	n
Kama	34.5	58
Rosa	31.5	53
Canadian	33.9	57
Total	100.0	168

Kama (57/116)

Table:

Category	%	n
Kama	49.1	57
Rosa	1.7	2
Canadian	49.1	57
Total	69.0	116

Rosa (51/52)

Table:

Category	%	n
Kama	1.9	1
Rosa	98.1	51
Canadian	0.0	0
Total	31.0	52

Canadian (50/54)

Table:

Category	%	n
Kama	7.4	4
Rosa	0.0	0
Canadian	92.6	50
Total	32.1	54

Kama (53/62)

Table:

Category	%	n
Kama	85.5	53
Rosa	3.2	2
Canadian	11.3	7
Total	36.9	62

Kama (49/52)

Table:

Category	%	n
Kama	94.2	49
Rosa	3.8	2
Canadian	1.9	1
Total	31.0	52

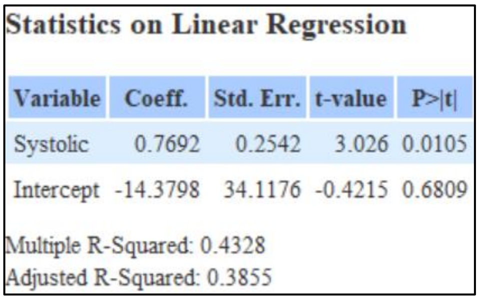
Canadian (6/10)

Table:

Category	%	n
Kama	40.0	4
Rosa	0.0	0
Canadian	60.0	6
Total	6.0	10

IF Length of kernel groove > 5.5755 THEN Group = 'Rosa' with probability 0.981.
IF Length of kernel groove <= 5.5755 AND Area <= 12.71 THEN Group = 'Canadian' with probability 0.926.
IF Length of kernel groove <= 5.5755 AND Area > 12.71 AND Asummetry coefficient <= 4.4135 THEN Group = 'Kama' with probability 0.942.

15

<table>
<tr><td>4a</td><td>Predictor variable: Systolic
Response varibale: Diastolic</td><td rowspan="5">15</td></tr>
<tr><td>4b</td><td>
<b>Statistics on Linear Regression</b>
<table>
<tr><th>Variable</th><th>Coeff.</th><th>Std. Err.</th><th>t-value</th><th>P&gt;|t|</th></tr>
<tr><td>Systolic</td><td>0.7692</td><td>0.2542</td><td>3.026</td><td>0.0105</td></tr>
<tr><td>Intercept</td><td>-14.3798</td><td>34.1176</td><td>-0.4215</td><td>0.6809</td></tr>
</table>
Multiple R-Squared: 0.4328
Adjusted R-Squared: 0.3855

Estimated_Diastolic = 0.7692×Systolic – 14.3798</td></tr>
<tr><td>4c</td><td>For every 1 mmHg increase in systolic pressure, the diastolic pressure is predicted to increase by 0.7692 mmHg.
The intercept value carries no meaning in this context.</td></tr>
<tr><td>4d</td><td>Estimated_Diastolic = 0.7692×120 – 14.3798 = 77.9 mmHg
(give full mark if only answer is given as students may use KNIME to generate the answer)</td></tr>
<tr><td>4e</td><td>R^2 is 0.4328 (not high) which suggests that the regression equation may not be useful for making estimation about the mean response variable.</td></tr>
<tr><td>5a</td><td>(i) 6
(ii) 2
(iii) Cluster_2, because distance to centroid_2 is smaller.</td><td rowspan="3">10</td></tr>
<tr><td>5b</td><td>(i) $\text{Centroid of Cluster } 1 = \frac{2+3+4}{3} = 3$
(ii) $\text{Centroid of Cluster } 2 = \frac{10+11+20+25+30+12}{6} = 18$</td></tr>
<tr><td>5c</td><td><i>k</i> = 3
This is because from the plot, the ‘elbow’ is indicated at <i>k</i>=3. It can be observed that by adding another cluster (say <i>k</i>=4) does not give much better modeling of the data in terms of reducing the WSSE. Hence, the best k is 3.
OR
This is because from the plot, the rate of decrease of WSSE sharply shifts when <i>k</i>=3 in the elbow plot. Hence, the best k is 3.</td></tr>
</table>

<table>
<tr>
<td>6a</td>
<td>
<table>
<tr><td>$P\left(<=50K\right)=16/20$</td><td>$P\left(>50K\right)=4/20$</td></tr>
<tr><td>$P\left(Education\le 9\,|\le 50K\right)=7/16$</td><td>$P\left(Education\le 9\,|> 50K\right)=1/4$</td></tr>
<tr><td>$P\left(Education> 9\,|\le 50K\right)=9/16$</td><td>$P\left(Education> 9\,|> 50K\right)=3/4$</td></tr>
<tr><td>$P\left(Marital\_Status=Married\,|\le 50K\right)=8/16$</td><td>$P\left(Marital\_Status=Married\,|> 50K\right)=2/4$</td></tr>
<tr><td>$P\left(Marital\_Status=Others\,|\le 50K\right)=8/16$</td><td>$P\left(Marital\_Status=Others\,|> 50K\right)=2/4$</td></tr>
</table>
</td>
<td rowspan="3">24</td>
</tr>
<tr>
<td>6b</td>
<td>
Let $\mathbf{x}=\left(Education>9\,years, Marital\_Status=Others\right)$

$$P\left(<=50K\,|\mathbf{x}\right)=\frac{P\left(\mathbf{x}|<=50K\right).P\left(<=50K\right)}{P\left(\mathbf{x}\right)}$$

$$=\frac{P\left(Education>9\,|<=50K\right).P\left(Marital\_Status=Others\,|<=50K\right).P\left(<=50K\right)}{P\left(\mathbf{x}\right)}$$

$$=\frac{\left(\frac{9}{16}\right)\left(\frac{8}{16}\right)\left(\frac{16}{20}\right)}{P\left(\mathbf{x}\right)}\approx\frac{0.225}{P\left(\mathbf{x}\right)}$$

$$P\left(>50K\,|\mathbf{x}\right)=\frac{P\left(\mathbf{x}|>50K\right).P\left(>50K\right)}{P\left(\mathbf{x}\right)}$$

$$=\frac{P\left(Education>9\,|>50K\right).P\left(Marital\_Status=Others\,|>50K\right).P\left(>50K\right)}{P\left(\mathbf{x}\right)}$$

$$=\frac{\left(\frac{3}{4}\right)\left(\frac{2}{4}\right)\left(\frac{4}{20}\right)}{P\left(\mathbf{x}\right)}\approx\frac{0.075}{P\left(\mathbf{x}\right)}$$

Since $\frac{0.225}{P\left(\mathbf{x}\right)}>\frac{0.075}{P\left(\mathbf{x}\right)}$, **x** is likely to earn <=50K per year.
</td>
</tr>
<tr>
<td>6c</td>
<td>
With independence assumption:

$$P\left(>9\text{years}\,|>50K\right)P\left(Others\,|>50K\right)=\frac{3}{4}.\frac{2}{4}=\frac{3}{8}$$
</td>
</tr>
</table>

7ai	$\text{Y}= -1312.2922 + 3.8165(\text{X1}) + 0.0444(\text{X2}) + 0.1525(\text{X3}) + 259.4698(\text{X4})$	3	
7a ii	The most influential variable is X4 as its coefficent is considerably larger than those of other predictor variables. In addition, its p-value also indicates that it has statistical significance.	2	
7a iii	The numerical meaning of the the coefficient of X4 implies that if 'Market Share' changes by a unit measure, the 'Sales' is expected to increase by 259.47 units.	2	
7a iv	$\text{Y}= -1312.2922 + 3.8165(40) + 0.0444(50000) + 0.1525(6000)$ $+ 259.4698(12)$ $= 5089.0054$ or $= 5090$ (corr. to 3 s.f.)	3	
7a v	The R^2 value is 0.896. This means that the linear regression model is able to explain 89.6% of the variation observed in the data. (It is unable to explain 10.4% of the variation in the data, which is thus attributed to error-based, random variation.)	2	
7f	The adjusted R^2 value adjusts the R^2 value for the number of predictor variables in the model. Its value increases only when a newly included predictor variable improves the model fit to the data by an amount more than what pure chance would do. So, the adjusted R^2 value may decrease when a newly included predictor variable does not improve the model fit by a sufficient amount.	2	
7g	Introduction of X5 in the model has increased both R^2 (0.915) and adjusted R^2 values (0.8926). In addition, the size of the coefficient of X5 is the largest of all the coefficients. So, X5 has large impact on the expected prediction of 'Sales' and this impact is also significant as suggested by the increase in R^2 and adjusted R^2 values.	2	16