SOLUTIONS

SINGAPORE POLYTECHNIC
2020 / 2021 Semester 2 EST

Module Name: Statistics and Analytics for Engineers
Course: DARE DASE DBEN DCPE DEB DEEE DME DMRO

Module Code: MS_SAE
Year: 2 FT
Page 1 of 2

MCQ

1.	2.	3.	4.	5.
c	b	b	d	b

<table>
<tr><th>Q1</th><th>Solution</th></tr>
<tr><td>a(i)</td><td><table><tr><td><b>Number of rows</b></td><td>1000</td></tr><tr><td><b>Number of columns</b></td><td>10</td></tr></table></td></tr>
<tr><td>a(ii)</td><td>2</td></tr>
<tr><td>b(i)</td><td>5</td></tr>
<tr><td>b(ii)</td><td>gender, last_new_job</td></tr>
<tr><td>c(i)</td><td><table><tr><td><b>(A)</b></td><td>47</td></tr><tr><td><b>(B)</b></td><td>60.171</td></tr></table><i><b>Accept (B) 60.2</b></i></td></tr>
<tr><td>c(ii)</td><td><table><tr><td><b>(A)</b></td><td>46</td></tr><tr><td><b>(B)</b></td><td>64.882</td></tr></table><i><b>Accept (B) 64.9</b></i></td></tr>
<tr><td>c(iii)</td><td>22</td></tr>
<tr><td>d</td><td><table><tr><td><b>Number of records after Step I</b></td><td>812</td></tr><tr><td><b>Number of records in ‘Very Experienced’</b></td><td>116</td></tr><tr><td><b>% of samples in ‘Very Experienced’</b></td><td>14.29 %</td></tr></table><i><b>Accept (B) 14.3%</b></i></td></tr>
<tr><td></td><td></td></tr>
</table>

<table>
<tr><th>Q2</th><th>Solution</th></tr>
<tr><td>a</td><td>k=2, since the data consists of only 2 groups (G and F)</td></tr>
<tr><td>b</td><td>The notes are relatively well separated using measurements on variance and skewness. Genuine notes on average appear to have higher values on these two measurements than fake notes. Measurements on entropy, on the other hand, do not appear to separate the notes well, as there is considerable overlap in their values between the two groups.</td></tr>
<tr><td>c</td><td>Different variables may be measured on different scales. Normalization is important so as not to let any variable with large numerical values to dominate the distance calculations.</td></tr>
<tr><td>d</td><td><table><tr><th></th><th>(I)</th><th>(II)</th></tr><tr><td><b>(A)</b></td><td>7.502</td><td>7.50</td></tr><tr><td><b>(B)</b></td><td>0.696</td><td>0.70</td></tr><tr><td><b>(C)</b></td><td>-9.583</td><td>-9.58</td></tr><tr><td><b>(D)</b></td><td>0</td><td>0</td></tr></table><i><b>Accept either (I) or (II)</b></i></td></tr>
<tr><td>e</td><td>$G = \left(\frac{0.696 + 0.909 + 1 + 0.982}{4}, \frac{0.739 + 1 + 0.721 + 0.36}{4}\right)$
$= (0.897, 0.705)$</td></tr>
<tr><td>f</td><td><table><tr><td>Standardized Variance</td><td>$\frac{2.20 - (-3.59)}{7.502}$</td><td>0.772</td></tr><tr><td>Standardized Skewness</td><td>$\frac{6.00 - (-9.583)}{17.421}$</td><td>0.894</td></tr><tr><td>$\boldsymbol{d}$(newnote, centroid$_G$)</td><td>$\sqrt{(0.772 - 0.897)^2 + (0.894 - 0.705)^2}$</td><td>0.227</td></tr><tr><td>$\boldsymbol{d}$(newnote, centroid$_F$)</td><td>$\sqrt{(0.772 - 0.208)^2 + (0.894 - 0.303)^2}$</td><td>0.817</td></tr></table>Since $\boldsymbol{d}$(newnote, centroid$_G$) &lt; $\boldsymbol{d}$(newnote, centroid$_F$) , the new note is likely to be genuine.
Note: Accept solution without standardisation.</td></tr>
<tr><td></td><td></td></tr>
</table>

SOLUTIONS

SINGAPORE POLYTECHNIC
2020 / 2021 Semester 2 EST

Module Name: Statistics and Analytics for Engineers
Course: DARE DASE DBEN DCPE DEB DEEE DME DMRO

Module Code: MS_SAE
Year: 2 FT
Page 2 of 2

Q3	Solution
a	bad_loans
b	Target variable is categorical
c	This is a classification task where the target 'bad loans' is a class label, that is categorical.
d	Using 80% of the data (i.e., 8000 records) to train the model, and 20% of the remaining data (i.e., 2000 records) to evaluate the accuracy of the model.
e	Gini index
f	First split occurs with the attribute "term (in months) <=48" or "term (in months) >48". Splitting with this attribute gives the lowest Gini index (or greatest gain) compared to other attributes (with reference to the unsplitted records.)
g	79.3% or 0.793 $\frac{1552+35}{2000}$ =0.7935
h(i)	68 68 borrowers were predicted to default payment of their loan but they did not.
h(ii)	345 345 borrowers who were predicted not to default payment of their loan, defaulted.
h(iii)	Having a high false negative error might result in the United Finance losing potential lender customers as the lenders might end up with bad debts due to default of payment by borrowers who were predicted not to default their payment. ***Accept other answers deemed logical***

<table>
<tr><th>Q4</th><th>Solution</th></tr>
<tr><td><b>a</b></td><td>House Price</td></tr>
<tr><td><b>b(i)</b></td><td>If there is a unit increase in bedroom, the house price will drop by $62,829</td></tr>
<tr><td><b>b(ii)</b></td><td>Estimated Price = -120018 + 45392*floors + 319.94*sqft_living - 35.1*sqft_above – 62829*bedrooms + 57093*condition - 0.672*sqft_lot</td></tr>
<tr><td><b>b(iii)</b></td><td>No.
Because the R-sq and R-sq(adj) values are close.</td></tr>
<tr><td><b>c</b></td><td>None
all coefficients are &lt; 5%</td></tr>
<tr><td><b>d(i)</b></td><td>48.45%</td></tr>
<tr><td><b>d(ii)</b></td><td>48.45% of the house price variation that is explained by the regression model</td></tr>
<tr><td><b>e</b></td><td>sqft_lot*condition,
p &gt; 0.05</td></tr>
<tr><td><b>f</b></td><td><table><tr><th>Steps</th><th>Name of the KNIME Node</th></tr><tr><td><b>Create the interaction term</b></td><td>Math Formula</td></tr><tr><td><b>To evaluate and score the model</b></td><td>Numeric Scorer</td></tr></table></td></tr>
<tr><td></td><td></td></tr>
</table>