VIT

Vellore Institute of Technology

School of Information Technology and Engineering

Winter Semester 2022-2023 - Fresher

Continuous Assessment Test – I

Programme Name & Branch: MCA

Course Name & code: Data Mining and Business Intelligence (ITA5007)

Class Number (s): 0528, 0296, 0530

Slot: C2+TC2

Faculty Name (s): Harshita PateL, Dr. Ephzibah E.P., Jagadeesan S.

Exam Duration: 90 Min. **Maximum Marks: 50**

<table>
<tr><th>Q.No.</th><th>Question</th><th>Max Marks</th></tr>
<tr><td>1.</td><td>There is a strong linkage between statistical data analysis and data mining. Some people think of data mining as an automated and scalable method for statistical data analysis. Do you agree or disagree with this perception? Present one statistical analysis method that can be automated and/or scaled up nicely by integration with the present data mining methodology.</td><td>10</td></tr>
<tr><td>2.</td><td>Briefly outline how to compute the dissimilarity between objects described by the following:
(a) Nominal attributes
(b) Asymmetric binary attributes
(c) Numeric attributes
(d) Term-frequency vectors</td><td>10</td></tr>
<tr><td>3.</td><td>Use these methods to normalize the following group of data:
200, 300, 400, 600,1000
(a) min-max normalization by setting min = 0 and max = 1
(b) z-score normalization
(c) normalization by decimal scaling</td><td>10</td></tr>
<tr><td>4.</td><td>Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:
<table>
<tr><td><i>age</i></td><td>23</td><td>23</td><td>27</td><td>27</td><td>39</td><td>41</td><td>47</td><td>49</td><td>50</td></tr>
<tr><td><i>%fat</i></td><td>9.5</td><td>26.5</td><td>7.8</td><td>17.8</td><td>31.4</td><td>25.9</td><td>27.4</td><td>27.2</td><td>31.2</td></tr>
</table>
<table>
<tr><td><i>age</i></td><td>52</td><td>54</td><td>54</td><td>56</td><td>57</td><td>58</td><td>58</td><td>60</td><td>61</td></tr>
<tr><td><i>%fat</i></td><td>34.6</td><td>42.5</td><td>28.8</td><td>33.4</td><td>30.2</td><td>34.1</td><td>32.9</td><td>41.2</td><td>35.7</td></tr>
</table>
(a) Calculate the mean, median, and standard deviation of age and %fat.
(b) Draw the boxplots for age and %fat.
(c) Draw a scatter plot and a q-q plot based on these two variables.</td><td>10</td></tr>
<tr><td>5.</td><td>Consider the following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
(a) Use smoothing by bin means to smooth these data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.
(b) How might you determine outliers in the data?
(c) What other methods are there for data smoothing?</td><td>10</td></tr>
</table>