

POKHARA UNIVERSITY

Level: Masters

Semester- Spring/Summer

Year: 2021

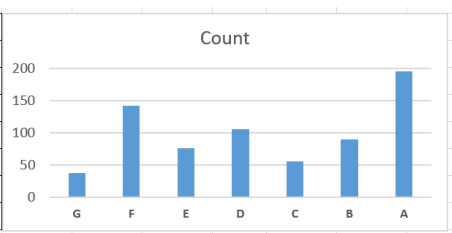
Programme: M.Sc. Computer Science / MCIS

Full Marks: 100

Course: Data Mining and Data Warehousing

Time: 4 hrs

Candidates are required to give their answers in their own words as far as practicable. The figures in the margins indicate full marks.
Attempt all the questions

Q.N.		Marks																																												
1.	(a) ‘Visualization/Interpretation’ is one of the various steps in Data Mining, why it is so important from the perspective of knowledge discovery through data-processing? (b) Data mining is defined as a multi-disciplinary subject, as it combines multiple technology concepts of Computer Science topics. List out at least four different such topics with brief explanation.	[6] [6]																																												
2.	What is datacube in data warehousing? What additional features and processes that datacube can provide? Explain briefly with appropriate example having 4-D cuboid and lattice structure.	[3+3+4] = 10																																												
3.	Consider below data about customer complaint on e-commerce portal and their service. Draw a free-hand sketch of Pareto diagram with clearly showing necessary calculations. <table><tr><th>Code</th><th>Complaint Types</th><th>Count</th></tr><tr><td>G</td><td>Unexpected fees</td><td>37</td></tr><tr><td>F</td><td>Unresponsive customer support</td><td>142</td></tr><tr><td>E</td><td>Lack of security</td><td>76</td></tr><tr><td>D</td><td>Poor display on mobile</td><td>105</td></tr><tr><td>C</td><td>Wrong product delivery</td><td>55</td></tr><tr><td>B</td><td>Product different than display</td><td>90</td></tr><tr><td>A</td><td>Slow website speed</td><td>195</td></tr></table> 	Code	Complaint Types	Count	G	Unexpected fees	37	F	Unresponsive customer support	142	E	Lack of security	76	D	Poor display on mobile	105	C	Wrong product delivery	55	B	Product different than display	90	A	Slow website speed	195	[8]																				
Code	Complaint Types	Count																																												
G	Unexpected fees	37																																												
F	Unresponsive customer support	142																																												
E	Lack of security	76																																												
D	Poor display on mobile	105																																												
C	Wrong product delivery	55																																												
B	Product different than display	90																																												
A	Slow website speed	195																																												
4.	Here is the sample transaction data records of a food and snacks serving stall on a particular morning of a day. Assume, the stall transaction record considers the each single bill record only, which might consists of either group of customers purchase or single customer purchase. Considering minimum support 30% and minimum confidence of 50%, answer the followings: <table><tr><th>Code</th><th>Item</th><th>Trx.</th><th>Item List</th></tr><tr><td>B</td><td>Bread</td><td>T1</td><td>M,B,O,T</td></tr><tr><td>C</td><td>Coffee</td><td>T2</td><td>S,T,M</td></tr><tr><td>M</td><td>Momo</td><td>T3</td><td>M,P,S,T</td></tr><tr><td>N</td><td>Noodle</td><td>T4</td><td>O,B,C,M,N</td></tr><tr><td>O</td><td>Omelette</td><td>T5</td><td>B,C,O,N</td></tr><tr><td>P</td><td>Pizza</td><td>T6</td><td>M,N,S,P,T</td></tr><tr><td>S</td><td>Samosa</td><td>T7</td><td>B,C,O,S,T</td></tr><tr><td>T</td><td>Tea</td><td>T8</td><td>S,N,M,T</td></tr><tr><td></td><td></td><td>T9</td><td>B,M,T,C,O</td></tr><tr><td></td><td></td><td>T10</td><td>P,C,S</td></tr></table> (a) Using Apriori algorithm for frequent pattern mining, identify the list of frequent items as per the s and c thresholds. (You need to show all the steps of calculations.) (b) Find out all strong association rules having at least two antecedent (i.e. $X \wedge Y \rightarrow Z$). (c) Will there be any additional rule generated if the s and c thresholds are lowered? If yes, write the new rule(s), if not why, justify.	Code	Item	Trx.	Item List	B	Bread	T1	M,B,O,T	C	Coffee	T2	S,T,M	M	Momo	T3	M,P,S,T	N	Noodle	T4	O,B,C,M,N	O	Omelette	T5	B,C,O,N	P	Pizza	T6	M,N,S,P,T	S	Samosa	T7	B,C,O,S,T	T	Tea	T8	S,N,M,T			T9	B,M,T,C,O			T10	P,C,S	4+4+4 = 12
Code	Item	Trx.	Item List																																											
B	Bread	T1	M,B,O,T																																											
C	Coffee	T2	S,T,M																																											
M	Momo	T3	M,P,S,T																																											
N	Noodle	T4	O,B,C,M,N																																											
O	Omelette	T5	B,C,O,N																																											
P	Pizza	T6	M,N,S,P,T																																											
S	Samosa	T7	B,C,O,S,T																																											
T	Tea	T8	S,N,M,T																																											
		T9	B,M,T,C,O																																											
		T10	P,C,S																																											
5.	Compare the followings: (any Three) a. Data normalization using Min-max vs. Z-score method b. Text mining vs. web mining c. Star schema vs. Galaxy schema d. Trend vs Seasonality component of time series data	[5*3] = 15																																												
6.	(a) What are the advantages of Naïve Bayes algorithm based classification? (b) Using the below data on insurance claims; find the followings: <table><tr><th>Vehicle Type</th><th>Cover</th><th>Years</th><th>Claim</th></tr><tr><td>SUV</td><td>Full</td><td>6</td><td>No</td></tr><tr><td>Compact</td><td>Partial</td><td>2</td><td>Yes</td></tr><tr><td>Compact</td><td>Full</td><td>4</td><td>Yes</td></tr><tr><td>SUV</td><td>Full</td><td>3</td><td>Yes</td></tr><tr><td>Compact</td><td>Partial</td><td>2</td><td>No</td></tr><tr><td>SUV</td><td>Partial</td><td>5</td><td>No</td></tr><tr><td>SUV</td><td>Full</td><td>3</td><td>Yes</td></tr><tr><td>SUV</td><td>Full</td><td>2</td><td>Yes</td></tr><tr><td>Compact</td><td>Full</td><td>4</td><td>No</td></tr><tr><td>Compact</td><td>Partial</td><td>5</td><td>No</td></tr></table> (i) Data column of ‘Year’ indicates the particular customer who has been continuing with us for their vehicle insurance, which is a continuous data type. Find out the best split value on year (like $\text{Years} < X$ and $\text{Years} \geq$) either using GINI or Entropy. (ii) Show your Naïve Bayes working based classification using the result of (i) for new data, where <i>Vehicle Type</i> is ‘SUV’, <i>Cover</i> is ‘Partial’ and <i>Years</i> is ‘2’.	Vehicle Type	Cover	Years	Claim	SUV	Full	6	No	Compact	Partial	2	Yes	Compact	Full	4	Yes	SUV	Full	3	Yes	Compact	Partial	2	No	SUV	Partial	5	No	SUV	Full	3	Yes	SUV	Full	2	Yes	Compact	Full	4	No	Compact	Partial	5	No	5 5+5
Vehicle Type	Cover	Years	Claim																																											
SUV	Full	6	No																																											
Compact	Partial	2	Yes																																											
Compact	Full	4	Yes																																											
SUV	Full	3	Yes																																											
Compact	Partial	2	No																																											
SUV	Partial	5	No																																											
SUV	Full	3	Yes																																											
SUV	Full	2	Yes																																											
Compact	Full	4	No																																											
Compact	Partial	5	No																																											
7.	What is ‘curse of dimensionality’? Why does subspace clustering become interesting problem in high dimensional data? Explain with appropriate example.	8																																												
8.	Explain the basics of DBSCAN algorithm and compare with any portioning clustering method.	8																																												
9.	Write short notes on the following (any Three): a. F_1 and F_α score b. Bootstrap method for model validation c. Outlier data identification using Box plot d. Ensemble method for classification	[3*4] = 12																																												