Reg.No.			

Manipal Institute of Technology

(Constituent Institute of Manipal University)

SIXTH SEMESTER B.TECH END SEMESTER EXAMINATIONS – MAY 2015
SUBJECT: INTRODUCTION TO DATA ANALYTICS- (MCA- 451)
18/05/2014

TIME: 3 HOURS **MAX.MARKS: 50**

Instructions to Candidates:

- **Answer any 5 FULL questions.**
- **All questions carry equal marks.**
- **Missing data may be suitably assumed.**

1A. Explain the different types of data analysis tasks with suitable examples.

1B. What are the different sources of data for analytics?

1C. How is the role of a Subject Matter expert different from the role of an IT expert in a data analytics project?

[5+3+2]

2A. Suppose that the data for analysis includes the attribute age. The age values of the data tuples are :

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.

i. Compute the 5 number summary.
ii. Clean the data by finding and eliminating outliers if any.
iii. Draw a box plot for the cleaned data.
iv. Use smoothing by bin means to smooth data using bins of depth size 3.

2B. Consider a data set with the values 250,370, 420, 605, 1100. Perform Data transformation on each of the above values with:

i. the min-max normalization method by setting min = 0 and max = 5
ii. the decimal scaling method

2C. How are missing values in a data set cleaned while preparing data for analysis?

[5+3+2]

3A Suppose a hospital tested the age and % of body fat for 10 randomly selected adults with the following result.

Age	23	23	27	27	39	41	47	49	50	52
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2	34.6

i. Visualize the data using a scatter plot.

ii. The Standard deviation of Age is 11.75 and standard deviation of %fat is 9.20. Calculate the correlation coefficient and determine if the two variables – age & % fat are positively or negatively correlated.

3B. An insurance company wanted to understand the time to process an insurance claim. They timed a random sample of 47 claims and determined that it took on average 25 minutes per claim and the standard deviation was calculated to be 3. With a confidence level of 95% (Z_c= 1.96), what is the confidence interval?

3C. How is a Contingency table different from a Summary table?

[5+3+2]

4A. A candidate rule has been extracted using the associative rule method from the table below: {Exhaustion = None, Stuffy node = Severe} => {Diagnosis = cold}

Calculate the support, confidence, and lift for this rule.

Patient id	Fever	Head-aches	General aches	Weak-ness	Exha-ustion	Stuffy nose	Sneezing	Sore throat	Chest disco-mfort	Diagn-osis
1326	None	Mild	None	None	None	Mild	Severe	Severe	Mild	Cold
398	Severe	Severe	Severe	Severe	Severe	None	None	Severe	Severe	Flu
6377	Severe	Severe	Mild	Severe	Severe	Severe	None	Severe	Severe	Flu
1234	None	None	None	Mild	None	Severe	None	Mild	Mild	Cold
2662	Severe	Severe	Mild	Severe	Severe	Severe	None	Severe	Severe	Flu
9477	None	None	None	Mild	None	Severe	Severe	Severe	None	Cold
7286	Severe	Severe	Severe	Severe	Severe	None	None	None	Severe	Flu
1732	None	None	None	None	None	Severe	Severe	None	Mild	Cold
1082	None	Mild	Mild	None	None	Severe	Severe	Severe	Severe	Cold
1429	Severe	Severe	Severe	Mild	Mild	None	Severe	None	Severe	Flu
14455	None	None	None	Mild	None	Severe	Mild	Severe	None	Cold
524	Severe	Mild	Severe	Mild	Severe	None	Severe	None	Mild	Flu
1542	None	None	Mild	Mild	None	Severe	Severe	Severe	None	Cold
8775	Severe	Severe	Severe	Severe	Mild	None	Severe	Severe	Severe	Flu
1615	Mild	None	None	Mild	None	Severe	None	Severe	Mild	Cold
1132	None	None	None	None	None	Severe	Severe	Severe	Severe	Cold
4522	Severe	Mild	Severe	Mild	Mild	None	None	None	Severe	Flu

4B. Differentiate between Single Link, Complete Link and Average Link clustering with a neat diagram.

4C. Why are the measures of support & confidence insufficient to assess the quality of the Association rule?

[5+3+2]

5A. Suppose the attributes of the colour of the car, type of car and its origin are recorded. The class label is the variable indicating whether the car was stolen – 'Yes' or 'No'. Predict the class label "Stolen" for a car with attributes – colour – red, type – SUV and origin – Domestic using the Naive Bayesian method

Colour	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

5B. Differentiate between the following, with suitable examples.

i. Supervised vs. unsupervised learning
ii. Eager vs. lazy learners
iii. Information Gain vs. Gain ratio

5C. Why is domain expertise important for a neural network based learning algorithm?

[5+3+2]

6A. Given the following dataset, with two predictor variables X1, X2 and response variable with class label '+' or '-'. Use the k-nearest neighbours algorithm to classify instance 7 & 8 for k=1 and k=3. Use majority voting and Manhattan distance for the calculation.

Instance	X1	X2	CLASS
1	0.25	0.25	+
2	0.25	0.75	+
3	0.50	0.25	-
4	0.50	0.75	-
5	0.75	0.50	-
6	0.75	1.00	+
7	**0.25**	**0.55**	**?**
8	**0.75**	**0.80**	**?**

6B. A classification prediction model was built using a training set of examples. A separate test set of 20 examples is used to test the model and the results are available in the table below. Calculate the model's accuracy measures:

i. Concordance
ii. Error rate
iii. Sensitivity
iv. Specificity

Observation	Actual	Predicted
1	0	0
2	1	1
3	1	1
4	0	0
5	0	0
6	1	0
7	0	0
8	0	0
9	1	1
10	1	1
11	1	1
12	0	1
13	0	0
14	1	1
15	0	0
16	1	1
17	0	0
18	1	1
19	0	1
20	0	0

6C. What strategy can be adopted to separate test data set from training data set for the purpose of classification or prediction?

[5+3+2]

----------*----------