Reg. No.									

INSPIRED BY LIFE

Manipal Institute of Technology

(A Constituent Institute of Manipal University)

VI SEMESTER B.Tech.
END SEMESTER EXAMINATION – MAY 2016

SUBJECT: INTRODUCTION TO DATA ANALYTICS [MCA 451]

16-05-2016

Time : 3 hours

Max. Marks : 50

Instructions to Candidates

1. Answer ANY FIVE FULL questions.
2. Missing data may be suitably assumed.

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

1B What is the role and responsibilities of a subject matter expert in data analysis projects?

1C Distinguish between dichotomous and nominal data variables using appropriate examples.

(5 + 3 + 2)

2A A training dataset consists of the following attributes and class label

Name	Age	Gender	Blood group	Weight (kg)	Height (m)	Systolic blood pressure	Diastolic blood pressure	Temperature (°F)	Diabetes
P. Lee	35	Female	A Rh^{+}	50	1.52	68	112	98.7	0
R. Jones	52	Male	O Rh^{-}	115	1.77	110	154	98.5	1
J. Smith	45	Male	O Rh^{+}	96	1.83	88	136	98.8	0
A. Patel	70	Female	O Rh^{-}	41	1.55	76	125	98.6	0
M. Owen	24	Male	A Rh^{-}	79	1.82	65	105	98.7	0
S. Green	43	Male	O Rh^{-}	109	1.89	114	159	98.9	1
N. Cook	68	Male	A Rh^{+}	73	1.76	108	136	99.0	0
W. Hands	77	Female	O Rh^{-}	104	1.71	107	145	98.3	1
P. Rice$^{'}$	45	Female	O Rh^{+}	64	1.74	101	132	98.6	0
F. Marsh	28	Male	O Rh^{+}	136	1.78	121	165	98.7	1

i. Create a new attribute "NormWeight" by normalizing the "Weight (kg)" attribute into the range of 0 to 1.

ii. Create a new attribute "AgeBins" by binning the Age attribute into 3 categories:

young (<30), middleAged (>= 30 and < 45) and old (>= 45 and above).

iii Create an aggregated column called BMI based on the formula

$$\mathbf{BMI} = \frac{\mathbf{Weight}\ (\mathrm{kg})}{\mathbf{Height}\ (\mathrm{m})^2}$$

iv. Segment the original data set into 2 data sets based on the variable Gender.

2B An insurance company wanted to understand the time to process an insurance claim. They timed a random sample of 45 claims and determined that it took on average 28 minutes per claim and the standard deviation was calculated to be 3.With a confidence level of 95% (Zc= 1.96), what is the confidence interval?.

2C How is a Contingency table different from a Summary table? Give examples.

(5 + 3 + 2)

3A Consider the transactional data set given below. Let minimum support be 60 %. Find all the frequent item sets only, using the Apriori algorithm.

TransactionID	**Items purchased**
100	Bread, Cheese, Eggs, Juice
200	Bread, Cheese, Juice
300	Bread, Milk, Yogurt
400	Bread, Juice, Milk
500	Cheese, Juice, Milk

3B What is the need for performing Correlation Analysis on association rules? Illustrate with an example.

3C How does the partition algorithm improve on the efficiency of the Apriori algorithm?

(5 + 3 + 2)

4A Consider the following distance matrix and perform agglomerative clustering on the 5 data points. Visualize using a dendrogram.

	p1	p2	p3	p4	p5
p1	0	0.10	0.41	0.55	0.35
p2	0.10	0	0.64	0.47	0.98
p3	0.41	0.64	0	0.44	0.85
p4	0.55	0.47	0.44	0	0.76
p5	0.35	0.98	0.85	0.76	0

4B Given two data points X= (20, 3, 40, 15) and Y= (14, 0 , 46, 8) .Represent them as a distance matrix using

i. Euclidean distance between the data points

ii. Manhattan distance between the data points.

iii. Minkowski distance between the data points using q = 3.

4C What are the disadvantages of the k-means clustering technique?

(5 + 3 + 2)

5A The following table shows the relationship between the amount of fertilizer used and the Height of a plant.

i. Calculate a simple linear regression equation using Fertilizer as the descriptor and Height as the response.

ii. Predict the height when fertilizer is 9.5.

Fertilizer	10	5	12	18	14	7	15	13	6	8	9	11	16	20	17
Height	0.7	0.4	0.8	1.4	1.1	0.6	1.3	1.1	0.6	0.7	0.7	0.9	1.3	1.5	1.3

5B Differentiate between the following, with suitable examples.

i. Classification tree vs. Regression tree

ii. Eager vs. lazy learners

iii. Sensitivity vs. Specificity

5C How do hyper planes perform classification in the Support vector machine (SVM) classifier?

(5 + 3 + 2)

6A Consider the following data set for a binary class problem. Calculate the information gain when splitting on attribute A and on attribute B. Which attribute would be selected for the root of the decision tree?

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	–
T	T	+
F	F	–
F	F	–
F	F	–
T	T	–
T	F	–

6B What strategies could be adopted for separation of test and training set for classifiers?

6C Describe any measure which can indicate the accuracy of prediction algorithms.

(5 + 3 + 2)

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