Reg. No.									

Manipal Institute of Technology

(A Constituent Institute of Manipal University)

VI SEMESTER B.Tech
MAKEUP EXAMINATION – JULY 2016
SUBJECT: INTRODUCTION TO DATA ANALYTICS [MCA 451]

08 -07-2016

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 stages in a Data Analytics project.

1B What are the different types of deliverables in a data analytics project?

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

(5 + 3 + 2)

2A Consider the table provided below, which records a series of retail transactions monitored by the main office of a computer store.

i. Generate a contingency table for the variables "Store" and "Product category".

ii. Generate a summary table, grouping by "Product category" which shows a count of the number of observations and the sum of the Profit ($) for each row.

iii. Create a histogram of "Sales Price ($)" using the following intervals: 0 to less than 250, 250 to less than 500, 500 to less than 750, 750 to less than 1000.

iv. Create a scatterplot showing Sales price ($) against Profit ($).

Customer	Store	Product category	Product description	Sale price ($)	Profit ($)
B.March	New York, NY	Laptop	DR2984	950	190
B.March	New York, NY	Printer	FW288	350	105
B.March	New York, NY	Scanner	BW9338	400	100
J.Bain	New York, NY	Scanner	BW9443	500	125
T.Goss	Washington, DC	Printer	FW199	200	60
T.Goss	Washington, DC	Scanner	BW39339	550	140
L.Nye	New York, NY	Desktop	LR21	600	60
L.Nye	New York, NY	Printer	FW299	300	90
S.Cann	Washington, DC	Desktop	LR21	600	60
E.Sims	Washington, DC	Laptop	DR2983	700	140
P.Judd	New York, NY	Desktop	LR22	700	70
P.Judd	New York, NY	Scanner	FJ3999	200	50
G.Hinton	Washington, DC	Laptop	DR2983	700	140
G.Hinton	Washington, DC	Desktop	LR21	600	60
G.Hinton	Washington, DC	Printer	FW288	350	105
G.Hinton	Washington, DC	Scanner	BW9443	500	125
H.Fu	New York, NY	Desktop	ZX88	450	45
H.Taylor	New York, NY	Scanner	BW9338	400	100

2B What are the different methods we can adopt to clean missing data?

2C Differentiate between nominal and ordinal data variables.

(5 + 3 + 2)

3A A producer of magnets wishes to understand whether there is a difference between four suppliers (A, B, C, and D) of alloys used in the production of the magnets. Magnets from the four suppliers are randomly selected and the magnets are recorded as either satisfactory or not satisfactory as shown in contingency table below. With a 95% confidence limit and using this information:

i. Specify the null and alternative hypothesis

ii. Calculate chi-square and determine whether the company can make the claim.

	Satisfactory	Not satisfactory	Total
Supplier A	28	2	30
Supplier B	27	3	30
Supplier C	29	1	30
Supplier D	26	4	30
Total	110	10	120

3B Calculate the following statistics: Mode, Median and Mean for the variable Age. The data values for Age are 35, 52, 45, 70, 24, 43, 68, 77, 45 and 28. Comment on the skewness of the data.

3C What does the z-score represent? How can it be used to measure standard error calculations during sampling?

(5 + 3 + 2)

4A Consider the following distance matrix and perform agglomerative clustering on the 6 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= (22, 3, 40, 12) and Y= (24, 0 , 46, 8)

i. Represent data as a data matrix.

ii. Represent data as a distance matrix using Euclidean distance.

iii. Represent data as a distance matrix using Manhattan distance.

4C While performing clustering on a data set, how do we represent the similarity between two binary variables? Explain with an example.

(5 + 3 + 2)

5A A company producing widgets needs to identify if the widgets they are producing are 'good' or 'bad'. The training data set measures three numeric properties P1, P2 and P3 for each widget. The training samples are as follows:

P1	P2	P3	Result
0	0.2	0.8	Good
9.2	0.7	1.5	Bad
4.9	0.1	2.9	Good
2.7	5.3	6.2	Bad
2.4	0	3.7	Good

Explain how the three-nearest neighbor algorithm would classify the following new data - P1= 6.3, P2= 5.1, P3 = 0.4. Use any distance measure of your choice.

5B What are the advantages of using the Neural networks for classification or prediction?

5C Differentiate between lazy learning and eager learning techniques. Give examples.

(5 + 3 + 2)

6A The following table shows the midterm and final exam grades obtained for students in the Data Analytics course.

Midterm exam	50	81	74	94	86	59	83	65	33	88	81	72
Final exam	63	77	78	90	75	49	79	77	52	74	90	84

i. Use the method of least squares to find an equation for the prediction of a student's final exam grade based on the student's midterm grade in the course.

ii. Predict the final exam grade of a student who scored 80 marks in the mid-term exam.

6B From the Confusion Matrix provided below, compute the following

i. Accuracy of the classifier
ii. Sensitivity
iii. Specificity

classes	buy_computer = yes	buy_computer = no	total
buy_computer = yes	6954	46	7000
buy_computer = no	412	2588	3000
total	7366	2634	10000

6C What is the need for the Laplacian correction in the Naïve Bayesian classification method?

(5 + 3 + 2)

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