Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

VI SEMESTER B.Tech. END SEMESTER EXAMINATIONS, MAY 2017

SUBJECT: INTRODUCTION TO DATA ANALYTICS [MCA 3282]

REVISED CREDIT SYSTEM

(03/05/2017)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer **ANY FIVE FULL** questions.
- Missing data may be suitable assumed.

1A. Explain with a neat diagram, the various data analysis tasks and methods. **5**

1B. What are the different sources of data for analysis? **3**

1C. What is the need for data segmentation? **2**

2A. Suppose that the data for analysis includes the attribute age. The age values of the data tuples are **5**

13,15,19,16,16,22,22,20,20,22,35,35,35,35,33,30,33,25,25,25,25,36,40,45,46,52,70.

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.

v. Visualize the binned data using a bar chart.

2B. A producer of magnets wishes to understand whether there is a difference between four suppliers (A, B, C, 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 the table below. Use the chi-square test with a confidence of 95% to determine if there is a relationship between the two variables. **3**

	Satisfactory	Not satisfactory
Supplier A	28	2
Supplier B	27	3
Supplier C	29	1
Suppler D	26	4

2C. Differentiate between Type I Error and Type II Error in a hypothesis test. 2

3A. Consider the following data representing the vital health statistics of 5 patients. 5

i. Convert it to a distance matrix using any distance measure of your choice.

ii. Perform agglomerative clustering on the 5 data points.

iii. Visualize using a dendrogram.

	Percentage body fat	Weight	Height	Chest	Abdomen
A	12.3	154.25	67.75	93.1	85.2
B	31.6	217	70	113.3	111.2
C	22.2	177.75	68.5	102	95
D	14.1	176	73	96.7	86.5
E	23.6	197	73.25	103.6	99.8

3B. An association rule has been extracted using association rule mining, from the table of patient records shown below. 3

RULE: If Exhaustion=None AND Stuffy nose= Severe THEN Diagnosis = cold

Calculate the support, confidence and lift for the rule.

Table of patient records

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

3C. How does the partition algorithm improve on the efficiency of the Apriori algorithm? 2

4A. The following table shows the relationship between the amount of fertilizer used and the height of a plant. 5

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

II. Predict the height when fertilizer is 10.3.

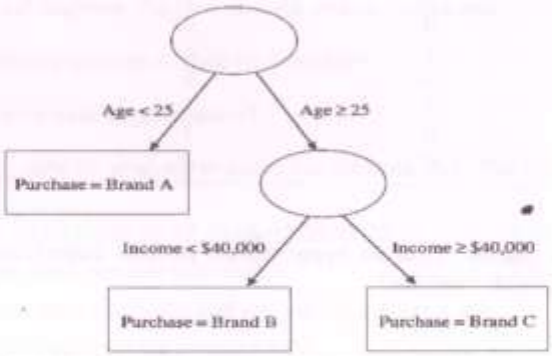
III. Visualize using a scatter plot.

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

4B. A classification tree has been built to predict the brand of printer a customer would purchase with a computer. 3

i. Write down all the classifier rules for the classification tree.

ii. For a customer whose age is 32 and Income is Rs. 35,000 , which brand of printer is he likely to buy?



4C. What does the graphical model of a Bayesian Belief network represent? 2

5A. Explain the architecture of multilayer feed forward neural network with a neat diagram. 5

5B. A classification prediction model was built using a training set of. 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: 3

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

5C. What are hyper planes? How do hyper planes perform classification in the Support vector machine (SVM) classifier? 2