Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

VI SEMESTER B.TECH. (INFORMATION TECHNOLOGY / COMPUTER AND COMMUNICATION ENGINEERING)

END SEMESTER EXAMINATIONS, APRIL 2017

SUBJECT: PROGRAM ELECTIVE II- BIG DATA ANALYTICS (ICT 4005)

(REVISED CREDIT SYSTEM)

(27/04/2017)

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

Instructions to candidates:

- Answer ALL the questions
- Missing data if any, may be suitably assumed.

1A.	Represent the logistic regression model and describe how logistic regression can be used as a classifier. How is ROC curve used in this model?	5
1B.	What is the method used to choose the value of k in kmeans clustering algorithms? Write an R code for the same.	3
1C.	What is Big Data? "Hadoop is a Big data technology" Justify.	2
2A.	For the dataset given in Table Q.2A, write a Map Reduce code to compute the number of flights from every Airline.	5
2B.	For successful analytical project which roles are significant and what are their responsibilities?	3
2C.	Consider the dataset given in Table Q.2A. Write a hive script to compute the number of direct flights. Hive script should contain creation of table, loading data from local file system and query.	2
3A.	Explain power of hypothesis testing and how sample size affects hypothesis testing. When can we apply student's t test? Check whether data1 and data2 have difference in means or not using Student's t test for the dataset given in Table Q.3A.	5
3B.	When is time series considered as stationary? Differentiate Auto-regressive model and Moving average models for time series.	3
3C.	What are the data structures used in data analytics? Mention the characteristics along with an example for each.	2
4A.	For successful text analysis, what are the different ways of collecting raw data and representing the data for processing? What is the significance of TF-IDF score? Explain with its generalized equation.	5

4B. Consider the dataset given in Table Q.2A. 3

i) Write a pig script to compute the number of flights from a given source.

ii) Write Pig script to compute on an average how many flights fly from each source which are run by another carrier.

4C. What is in-database analytics? What is advantage of using in-database analytics? 2

5A. Explain data analytical life cycle with a neat diagram and key activities in each phase. 5

5B. Write SQL query to illustrate how moving averages can be implemented using window functions. 3

5C. Write a R code for applying decision tree algorithm for building a model and predicting grade of a student in a subject given the independent attributes as id, marks1, marks2, assgnmarks. 2

Table Q.2A

Description and sample data	
Airline	2-letter (IATA) or 3-letter (ICAO) code of the airline.
Airline ID	Unique OpenFlights identifier for airline.
Source airport	3-letter (IATA) or 4-letter (ICAO) code of the source airport.
Source airport ID	Unique OpenFlights identifier for source airport
Destination airport	3-letter (IATA) or 4-letter (ICAO) code of the destination airport.
Destination airport ID	Unique OpenFlights identifier for destination airport
Codeshare	"Y" if this flight is a codeshare (that is, not operated by *Airline*, but another carrier), empty otherwise.
Stops	Number of stops on this flight ("0" for direct)
Equipment	3-letter codes for plane type(s) generally used on this flight, separated by spaces

Sample entries

```
BA,1355,SIN,3316,LHR,507,,0,744 777
BA,1355,SIN,3316,MEL,3339,Y,0,744
TOM,5013,ACE,1055,BFS,465,,0,320
```

Table Q.3A

Sl	1	2	3	4	5	6	7	8	9	10	11
data1	3	3	3	12	15	16	17	19	23	24	32
data2	20	13	13	20	29	32	23	20	25	15	30

		Significance level					
Degrees of freedom	*Two-tailed test:* / *One-tailed test:*	10% / 5%	5% / 2.5%	2% / 1%	1% / 0.5%	0.2% / 0.1%	0.1% / 0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221

T distribution table : critical values of t

****