Reg. No.									

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

(A constituent unit of MAHE, Manipal)

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

END SEMESTER EXAMINATIONS, APRIL 2018

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

(REVISED CREDIT SYSTEM)

(24/04/2018)

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

Instructions to candidates:

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

1A.	Represent linear regression model with normally distributed error. Explain four tools and techniques that can be used to validate a fitted linear regression model.	5
1B.	Let x,y be the variable of the dataset, x={2.5, 3, 4, 7, 5, 1.8, 2, 4} y={1.5, 3, 4.2, 6, 2.3, 3, 5, 2.3}, consider second and sixth points as initial centroids and apply k-means algorithm to compute centroids after two iterations.	3
1C.	Consider the data given in Table Q.1C, Given a new example as W = F, X = T, Y = F. How should this example be classified using the Naive Bayes method? Show your computations.	2
2A.	Given medicine response data and tabulation for critical values in Table Q.2A, apply suitable hypothesis test to check if all the medicines have same effect or not?	5
2B.	Explain stemming, Bag of words, Topic modeling with respect to Text Analytics.	3
2C.	Write the steps for writing and executing Map Reduce program.	2
3A.	Explain each component of ARIMA model. Mention the use of Box-Jenkins methodology while using this model.	5
3B.	What are the characteristics and major categories of NoSQL? Explain.	3
3C.	Given the input file as product.csv containing (year, product, quantity), write a hive script to select products whose quantity is greater than 1000 and year is 2001.	2
4A.	Explain first two phases of data analytical life cycle with all the key activities involved in each.	5

4B. Consider the dataset with schema (id, BookName, Author, NumberofPages, ReadbyCount, LikedbyCount) 3

i) Write a pig script to display book and authors names which are liked by more than 100 readers

ii) Write Pig script to compute on an average read counts for the books written by each author.

4C. What are the key outputs for each of the stakeholders in data analytical project? 2

5A. What are the components to be configured in Driver class for Map Reduce? Consider the dataset with schema (id, BookName, Author, publisher, NumberofPages, ReadbyCount, LikedbyCount), write Map and Reduce functions for computing the average likes for each publisher. 5

5B. Explain typical analytical architecture. What challenges do data scientist face with this architecture for advanced analytics. 3

5C. What R command(s) would be used to remove null values from a dataset? Give example with simple data. 2

Table Q.1C

W	X	Y	C
T	T	T	T
T	F	T	F
T	F	F	F
F	T	T	F
F	F	F	T

Table Q.2A

M1	19	20	21	21	19
M2	18	19	17	20	19
M3	15	18	19	17	19

Critical values of F for the 0.05 significance level:

	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.97	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.10	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60

****