Reg. No.									

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

(A constituent unit of MAHE, Manipal)

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

MAKEUP EXAMINATIONS, JUNE 2018

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

(REVISED CREDIT SYSTEM)

(20/06/2018)

TIME: 3 HOURS MAX. MARKS: 50

Instructions to candidates:

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

1A.	Represent logistic regression model and write R code for building and testing logistic regression model. What are the metrics used to test the model?	5
1B.	Why term frequency alone can not be used as a measure for finding usefulness of the words? Explain with equation the inverse document frequency.	3
1C.	In the bank marketing example, the training set includes 2,000 instances. An additional 100 instances are included as the testing set. Construct the confusion matrix for a classifier on 100 clients to predict whether they would subscribe to the term deposit given, the 11 clients who subscribed to the term deposit, the model predicted 3 subscribed and 8 not subscribed. Similarly, of the 89 clients who did not subscribe to the term, the model predicted 2 subscribed and 87 not subscribed. Compute Precision and Specificity.	2
2A.	Consider the data given in Table Q.2A. and apply Wilcoxon rank sum hypothesis testing to check if both samples are same. Mention null and alternate hypothesis. Consider the critical value as 115.	5
2B.	How is goodness of the clusters measured? Explain	3
2C.	Explain the components of Pig Architecture. What are the execution modes possible in Pig?	2
3A.	Consider the load data having fields [id, borrowerName, Purpose, loanAmount, installmentsPaid] stored as loan.csv. Write Map Reduce functions to display average loan amount for the purpose where more than 10 installments are paid.	5
3B.	Explain CAP theorem and BASE compatibility with respect to NOSQL.	3
3C.	Explain hadoop architecture and its elements with neat diagram	2

4A.	Explain the significance and steps of model planning and model execution phases of data analytical life cycle with all the key activities involved in each.	5
4B.	Consider the dataset population.csv containing Year, Age, Gender (male=1 &female=2), Population count for particular age and gender i) Write a pig script to pull out all the result for 55 year old women. ii) Write Hive script to compute on an average age of male and female population for years greater than 1997.	3
4C.	Construct the root node of the decision tree for the data given in Table Q.4C.	2
5A.	Write necessary conditions required for a time series to be stationary with equations. Write R code for building and evaluating ARIMA model.	5
5B.	Explain typical analytical architecture. What challenges do data scientist face with this architecture for advanced analytics.	3
5C.	What R function is used to encode a vector as a category? Explain with an example	2

Table Q.2A

S1	2	3.6	2.6	2.6	7.3	3.4	14.9	6.6	2.3	2.0	6.8	8.5
S2	3.5	5.7	2.9	2.4	9.9	3.3	16.7	6.0	3.8	4.0	9.1	20.9

Table Q.4C

	yes	no
Sunny	2/9	3/5
Overcast	4/9	0/5
rainy	3/9	2/5

****