Reg. No.										

Manipal Institute of Technology
MANIPAL

(A constituent unit of MAHE, Manipal)

VII SEMESTER B. TECH. (COMPUTER SCIENCE AND ENGINEERING)
REGULAR EXAMINATIONS, November 2019
SUBJECT: ELECTIVE V – MACHINE LEARNING WITH BIG DATA [CRA 4007]
REVISED CREDIT SYSTEM
(26/11/2019)

Time: 3 Hours Max. Marks: 50

Instructions to Candidates:
• Answer **ALL** questions & missing data may be suitably assumed.

1.A. What is data preprocessing? What constitutes data preprocessing? Explain each component and subcomponent of data-preprocessing with the help of examples. 5M

1.B. List and explain the phases of CRISP-DM 3M

1.C. Compare KNIME with Spark MLib. 2M

2.A. Calculate the measure of location and spread for the following data 5M

A	35	42	78	25	60	50	42	78	81	87

2.B. Suppose the fraction of undergraduate students who use mobile is 15%, and the fraction of graduate students who use mobile is 23%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who uses mobile is a graduate student? 3M

2.C. In a decision tree, when to stop splitting a node? 2M

3.A. Write the pySpark code to achieve the following 5M
A. Read the data into the dataframe
B. Print the data types of all the columns
C. Print Summary Statistics
D. Print the number of columns
E. Print the number of rows

3.B. What is generalization? Why is generalization important in machine learning? How generalization and overfitting are related? 3M

3.C. What is the role of confusion matrix in machine learning? What do the diagonal elements of confusion matrix represent? 2M

4.A.	Explain Pre-pruning and Post-pruning giving examples for each. Which of these two pruning techniques gives better result and why?	5M
4.B.	What are the limitations to the holdout method?	2M
4.C.	Explain the working of cross validation method with a neat diagram.	3M
5.A.	Define Precision and Recall.	2M
5.B.	Explain the algorithm for clustering using K-Means.	3M
5.C.	Define Frequent itemset, Support(x) and confidence with respect to association rules. . Write all the steps of association process analysis.	5M