Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent unit of MAHE, Manipal

INSPIRED BY LIFE

VII SEMESTER B.TECH. (COMPUTER SCIENCE & ENGINEERING)

END SEM EXAMINATIONS, Nov/Dec 2018

SUBJECT: MACHINE LEARNING WITH BIG DATA [CRA- 4007]

REVISED CREDIT SYSTEM

(29/11/2018)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A.	Briefly explain the following methods used in graphic displays of basic statistical description of data: i. Boxplot ii. Scatter plot	**4M**
1B.	Explain the following data pre-processing methods: i) Feature Transformation ii. Principal component analysis	**4M**
1C.	With an example how correlation coefficient is used as a measures of dependence to describe relationship between variables	**2M**
2A.	Describe the process of constructing a decision tree. Also, Explain how a decision tree is used for classification.	**4M**
2B.	Briefly outline the different steps of naïve Bayesian classifier.	**4M**
2C.	With a diagram explain the general steps in building a classifier.	**2M**
3A.	Describe how kNN is used for classification.	**5M**
3B.	Distinguish between overfitting and under fitting. What is generalization? Describe how overfitting is related to generalization, and explain why overfitting should be avoided.	**5M**
4A.	Describe several ways to create and use the validation set to address overfitting.	**5M**
4B.	Describe how a confusion matrix can be used to evaluate a classifier. Illustrate with the following result of a binary classifier which classifies whether a given animal is	**3M**

mammal or not.

True Label	Yes	No	No	Yes	Yes	No	Yes	Yes	No	No
Predicted Label	No	No	No	Yes	Yes	No	No	Yes	No	Yes

4C. What is regression? Explain the difference between regression and classification, and name some applications of regression **2M**

5A. With necessary examples, discuss the usage of cluster analysis in data segmentation, classification of new data samples and anomaly detection **3M**

5B. Describe the steps in the k-means algorithm. Explain when to stop iterating when using k-means? How you address the sensitivity of final clusters of initial centroids? **4M**

5C. For the following transaction table, find the frequent 1-, 2- and 3- item sets with a minimum support of 60%. Generate an association rule, if any, from 3-itemsets with minimum confidence of 0.95. **3M**