Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL

A Constituent unit of MAHE, Manipal

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

MAKE UP EXAMINATIONS, Dec 2018

SUBJECT: MACHINE LEARNING WITH BIG DATA [CRA 4007]

REVISED CREDIT SYSTEM

(31/12/2018)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A. In real-world data, tuples with *missing values* for some attributes are a common occurrence. Describe various methods for handling this problem. **4M**

1B. Explain the different ways of measuring the central tendency of data. Show their relationship in symmetric, positively skewed and negatively skewed data with a diagram. **4M**

1C. With an example, explain how smoothening by bin -means helps in preprocessing. **2M**

2A. Explain different steps involed in building and applying a classification model. **5M**

2B. Explain how Naïve bayesian classifer is built and used for classification. **5M**

3A. Discuss the benefits of dimensionality reduction and explain how PCA helps in dimensionality reduction. **5M**

3B. Discuss overfitting in the context of decision tree models. Explain how can overfitting be addressed in decision trees with pre-pruning and post-pruning. **5M**

4A. What is a validation set? Explain how it relates to overfitting and model performance evaluation. **5M**

4B. How do you evaluate the performance of a classifier? Explain the different metrics available. Why some of them fail with class imbalance problem and how it can be addressed? **3M**

4C. Describe how linear regression works and explain how least squares is used in linear regression, **2M**

5A. Explain the goal of cluster analysis and discuss the different metrics used to measure similarity between samples. **4M**

5B. How you evaluate clustering results? By what means you choose the optimal value for k in k-means? **3M**

5C. With an example, list the basic steps followed in Association Analysis. Discuss the usage of association analysis in market basket analysis, recommendation systems and medical applications. **3M**