Q.P. Code: 40829

Time: 03 Hours **Marks: 80**

Note: 1. Question 1 is compulsory

2. Answer any three out of remaining questions.
3. Assume suitable data wherever required and justify the same.

Q1 a) Describe how logistic regression can be used as a classifier. **[5]**

b) Explain cross-validation for accuracy estimation. **[5]**

c) What is data journalism? **[5]**

d) What is data leakage with respect to big data? **[5]**

Q2 a) What is type I and type II errors in hypothesis testing? Is one always more serious than the other? Why? **[10]**

b) Describe the working of the Map-Reduce with an example. **[10]**

Q3 a) Explain Gaussian (normal) distribution with respect to pdf and cdf and its use in statistics. **[10]**

b) Explain time series mining with an appropriate example. **[10]**

Q4 a) You have collected a data of about ten thousand rows of tweet text. With help of text mining how you will create a tweet classification model that categorizes each of the tweets in three different classes. What could be the challenges while performing text mining to this context? **[10]**

b) Explain the process of content based RS with suitable example. **[10]**

Q5 a) Explain singular value decomposition (SVD) with an example. **[10]**

b) What infrastructure is most appropriate for Hadoop? Draw and describe Hadoop Ecosystem Architecture. **[10]**

Q6 a) Given $S = \begin{matrix} 1 & -1 \\ 0 & 1 \\ -1 & 0 \end{matrix}$ **[10]**

Find principal components.

b) Draw and describe the information visualization process. **[10]**

* * *