U.S.N [ ][ ][ ][ ][ ][ ][ ][ ][ ][ ]

P.E.S. College of Engineering, Mandya - 571 401
(An Autonomous Institution affiliated to VTU, Belagavi)
Eighth Semester, B.E. - Computer Science and Engineering
Semester End Examination; Aug. / Sep. - 2020
Data Science

Time: 3 hrs   Max. Marks: 100

Note: i) Answer TWO full questions, selecting ONE full question from UNIT - I and UNIT - II.
ii) Answer any THREE full questions, choosing from UNIT - III, UNIT - IV and UNIT - V.

UNIT - I

1 a. With a neat diagram, explain in brief the overview of the main phase of the data analytics life cycle.   10

b. Describe the key outputs for each of the main stakeholders of an analytics project and what they usually expect at the conclusion of the project?   10

OR

2 a. Explain the role of data exploration and variable selection in model planning phase.   10

b. List and explain common tools for the model building phase.   10

UNIT - II

3 a. Distinguish NOIR attribute types with example and show the operations they support.   10

b. With an example, explain NULL Hypothesis and alternative Hypothesis.   10

OR

4 a. With an 'R' code, explain the following :
   i) Data Frames
   ii) Lists
   iii) Factors   10

b. List and explain-'R' functions for visualizing a single variable.   10

UNIT - III

5 a. Explain the steps of k-means algorithm to find 'k' clusters.   10

b. Explain the evaluation of candidate rules in Apriori algorithm.   10

6 a. Explain some specific applications of k-means clustering.   10

b. Write an algorithm for Apriori.   10

UNIT - IV

7 a. Explain N-fold cross validation.   10

b. Discuss how the ROC curve can be used to determine an appropriate threshold value for a classifier?   10

8 a. Explain the common use cases of logistic regression model.   10

b. Write an ID3 Algorithm.   10

UNIT - V

9 a. Explain in detail the Map-Reduce paradigm to break a large task into smaller tasks, sum the tasks in parallel and consolidate the outputs of the individual tasks into the final output.   10

   b. With an example, explain Four major categories of NOSQL tools.   10

10 a. Explain HDFS to distribute data across a cluster to take the advantage of the parallel processing of Map-Reduce.   10

   b. Explain the following Hadoop ecosystem:
       i) Pig   10
       ii) Hive

* * * *