(2 ½ Hours) **[Total Marks: 75]**

N.B. 1) All questions are compulsory.
2) Figures to the right indicate marks.
3) Illustrations, in-depth answers and diagrams will be appreciated.
4) Mixing of sub-questions is not allowed.

Q. 1 Attempt All

(a) Select the correct alternative from the options given: **(10M)**

(i) Exploratory Data Analysis represents data in _____ format.
(a) Numerical (b) Character
(c) String (d) Graphical

(ii) _____ interviews are conducted by a trained interviewer in a non-structured and natural way with a small group.
(a) focus group (b) observation
(c) formal (d) informal

(iii) Imputation or removal of data are used during handling of _____ data.
(a) collected (b) Missing
(c) table (d) Duplicate

(iv) ____ is a query language used for traversing through an XML document.
(a) XML (b) TQML
(c) Xquery (d) Xpath

(v) _____ data have semantic tags.
(a) structured (b) unstructured
(c) semi structured (d) unorganised

(vi) In version control _____ is a mainline or unique line of the development which is not actually a branch.
(a) sub branch (b) trunk
(c) path (d) root

(vii) _____ service of cloud support services such as storage and network connectivity on demand.
(a) IaaS (b) PaaS
(c) SaaS (d) SaaN

(viii) AIC is suited over BIC when the model is _____.
(a) simple (b) complex
(c) large (d) Small

(ix) Lasso regression was introduced in order to improve the prediction _____ and interpretability.

(a) accuracy (b) values
(c) result (d) set

(x) _____ is the process of making prediction of the future based on present and past data.

(a) Trend (b) Seasonality
(c) forecasting (d) classification

(b) **Fill in the blanks by selecting from the pool of options:** **(5M)**
(aggregation, unstructured, discrete, disguised, supervised, personal, unsupervised, smoothing, structured, continuous)

(i) Apriori, K-means and K-medoids are the example of _____ learning algorithm.

(ii) _____ deals with removal of noise from data.

(iii) _____ data are not organized into special repositories.

(iv) In _____ observation the person who is being observed is unaware that he is being observed.

(v) Height and weight are the example of _____ data.

Q. 2 Attempt the following (Any THREE) **(15M)**

(a) What is data? Explain types of data.
(b) What is EDA? Explain methods to visualize data.
(c) What is data normalization? Illustrate any one type of data normalization technique with an example.
(d) Explain the difference between data and information.
(e) Describe any two types of observational methods used in data collection.
(f) Write a short note on data cleaning and data extraction.

Q. 3 Attempt the following (Any THREE) **(15M)**

(a) Discuss the 5 V's of data.
(b) What is MongoDB? State its features.
(c) How to create indexes in MongoDB? Give example.
(d) What is NoSQL? What are its features?
(e) Explain how you can read JSON file in R with the help of an example.
(f) Write a short note on AWS.

Q. 4 Attempt the following (Any THREE) **(15)**

(a) What are AIC, BIC? State their mathematical formula.
(b) Explain Forecasting. List the steps in forecasting.
(c) Write a short note on SVM.
(d) What is K-NN? Explain with the help of an example.
(e) Explain the filter method and forward selection method of data selection.
(f) Discuss the steps involved in implementing PCA on a 2-D Dataset.

Q. 5 Attempt the following (Any FIVE) **(15)**

(a) Explain the terms data, information and knowledge.
(b) Write a short note on Smoothing by means technique.
(c) How can you see data stored in MongoDB? Explain any two methods with example.
(d) Explain any 3 ways to do web scraping.
(e) Discuss the important characteristics of HBase.
(f) Give the formula for Information Gain and Entropy.
(g) Discuss Model, Train Data and Test Data.
(h) Discuss the Advantages of Dimensionality reduction.

****