Data Science and Big Data Analytics Laboratory Oral Question Bank

Allocation of problem statement to students is on random basis with help of chits. (No change in program once selected)

Every student will get questions for Orals on

1. Data Science questions based on theory, lab
2. Scala
3. Hadoop

Sample Oral Questions

1. Complete information of dataset used in given problem statement. Features of datasets, columns etc
2. Complete information of python libraries used in Data science. For example pandas, numpy, matplotlib, seaborn.
   a. What is use of these libraries
   b. What are different functions of these libraries
3. From where do we get the standard datasets
4. How do we fill missing values? What are techniques used for same? What are functions given by pandas to fill the missing values
5. How to handle the inappropriate data and inconsistencies in the preprocessing phase?
6. How to turn categorical variables into quantitative variables in Python
7. How to do data Normalization
8. How to handle outliers
9. What is skewness?
10. What is mean, median, minimum, maximum, standard deviation? How to handle it in python libraries
11. How to provide summary statistics of income grouped by the age groups.
12. Which libraries are used to display some basic statistical details like percentile, mean, standard deviation
13. What is Linear Regression Model? How to design it via python library
14. What is logistic regression? What is advantage of it?
15. What is confusion matrix? How to calculate TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall?
16. What is Simple Naïve Bayes classification algorithm
17. What is Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization, Term Frequency and Inverse Document Frequency.
18. What is histogram? How to draw histogram using seaborn library
19. What is box plot? How to implement it
20. What are attribute types? (e.g., numeric, nominal)
21. How to read text input using python libraries

22. Study of complete eco system of Hadoop. Learn each module of Hadoop ecosystem in details
    a. HDFS: Hadoop Distributed File System
    b. YARN: Yet Another Resource Negotiator
    c. MapReduce: Programming based Data Processing
    d. Spark: In-Memory data processing
    e. PIG, HIVE: Query based processing of data services
    f. HBase: NoSQL Database (Provides real-time reads and writes)
    g. Mahout, Spark MLLib: (Provides analytical tools) Machine Learning algorithm libraries
    h. Solar, Lucene: Searching and Indexing
23. How to do map-reduce programming? What is advantage of it
24. What is a scala programming? features of scala programming
25. What is scikit-learn library