Foundations of Data Science

Important Questions

Mid Exam 1

1. Calculate PCA for any matrix of your choice with the specified number of principal components.
2. Write a Python code to find PCA for the given dataset of your choice.
3. Differentiate the role, features, and importance of Big Data and Data Science.
4. Explain the probability distribution techniques.
5. Write a Python code to find Eigenvalues and Eigenvectors.
6. Calculate SVD for any matrix of your choice.
7. Write a Python code to find SVD for any dataset of your choice.
8. What are the data modeling stages involved?
9. Write Python code & Formula to calculate: Mean, Mode, Median, Variance, Standard Deviation, correlation and Coefficient of variation.
10. Differentiate the features of supervised machine learning & unsupervised learning. Give an example.
11. Apply row reduction to find the 'Echelon' form for the matrix of your choice. Write Python code for the above-mentioned technique.
12. Write a Python code to diagnolize the given matrix using Eigen decomposition.
13. Explain the classification model validation metrics in detail.
14. What is the difference between the validation metrics accuracy measure and the F1 Score in the classification techniques?
15. Student's t-test calculation for the dataset of your choice. How a student's t-test is used in classification. Write a Python code for the same.
16. Explain the following:
    a. Sample and Population
    b. Inferential Statistics and Descriptive Statistics
    c. Box Plot

d. IQR, Q1, Q2, Q3 calculation, and Python

e. Data Discretization

f. Confusion matrix

g. Difference between the precision and recall

17. Python code for the histogram, scatter, and box plot.
18. Write a Python code to find AUC.

Write a formula if required and write a Python code for the following:

19. Write a formula & Python code for the following exploratory data analytics.
    - Data cleaning - Missing value imputation
    - Data cleaning- Binning method for removing noisy data
    - Data transformation – Normalizing features
    - Data reduction – Data cube aggregation
    - Z-Score calculation
    - Missing value imputation
    - Techniques to handle categorical values
    - Techniques to handle outliers
    - Importance of correlation
    - Correlation
    - Multicollinearity
    - Feature engineering techniques
20. Find the outliers using Z-Score, Box Plot, and Scatter plot using Python code for the given dataset.

    Item_price = (10, 400, 23, 45, 345, 56, 67, 78, 123, 245, 90, 5000, 692)
21. Explain the different applications/scenarios of supervised machine learning algorithms.