

Data Science Questions and Answers

3-Mark Questions

Q: Data Science Tools and Techniques

A: Data Science tools include Python, R, SQL, Hadoop, and Tableau; techniques include data cleaning, data visualization, machine learning, and statistical analysis.

Q: Introduction to Data Science

A: Data Science is the field that uses tools, techniques, and algorithms to collect, process, analyze, and interpret large data sets to extract meaningful insights and support decision-making.

Q: Applications of Data Science

A: Applications of Data Science include healthcare (disease prediction), finance (fraud detection), e-commerce (recommendation systems), and business (customer analytics).

Q: Data Science Lifecycle

A: The Data Science lifecycle includes: data collection, data preparation, data analysis/modeling, evaluation, and deployment of results.

Q: Data Analysis and Data Analysis Process

A: Data Analysis is examining data to find patterns and insights. The process includes: data collection → cleaning → exploration → analysis → interpretation.

Q: Data Science Tools

A: Common Data Science tools are Python, R, SQL, Jupyter Notebook, Hadoop, Spark, and Tableau.

4-Mark Questions

Q: Machine Learning, Deep Learning – Definitions with Examples

A: Machine Learning: A method where computers learn patterns from data and make predictions without being explicitly programmed. Example: Predicting house prices. Deep Learning: A subset of ML that uses multi-layered neural networks to handle complex data like images, speech, and text. Example: Face recognition in smartphones.

Q: Data Science vs Machine Learning

A: Data Science: Broad field for collecting, cleaning, analyzing, and visualizing data. Machine Learning: Subset of Data Science that builds algorithms to learn from data and make predictions. Example: Data Science may analyze customer behavior, while ML predicts future purchases.

Q: Stages of Machine Learning

A: 1. Data Collection 2. Data Preparation 3. Model Training 4. Model Evaluation & Deployment

Q: Types of Machine Learning

A: 1. Supervised Learning – Uses labeled data (e.g., exam score prediction). 2. Unsupervised Learning – Finds hidden patterns (e.g., customer clustering). 3. Reinforcement Learning – Learns by trial and error with rewards (e.g., game AI).

Q: Dataset Anomalies

A: 1. Missing Values 2. Outliers 3. Duplicates 4. Inconsistencies

Q: Dataset Splitting: Training and Testing

A: Training Dataset: Used to train models. Testing Dataset: Used to check accuracy on unseen data. Example: 80% training, 20% testing.

Q: Overfitting and Underfitting

A: Overfitting: Model memorizes training data, performs poorly on new data. Example: Memorizing answers. Underfitting: Model too simple, fails to capture patterns. Example: Using a straight line for complex data.

Q: Python Libraries for Machine Learning

A: 1. Scikit-learn 2. TensorFlow 3. Keras 4. PyTorch

Q: NumPy Introduction

A: NumPy is a Python library for scientific computing. It supports arrays, matrices, and mathematical operations. Example: Matrix multiplication.

Q: Pandas (Data Cleaning)

A: Pandas is used for data manipulation and cleaning. Handles DataFrames, removes missing values, duplicates, and inconsistencies. Example: `df.dropna()` removes null values.

Q: SciPy Introduction

A: SciPy is a scientific computing library built on NumPy. It supports optimization, integration, linear algebra, and statistics. Example: Solving differential equations.

Q: Matplotlib Introduction

A: Matplotlib is a Python library for data visualization. It creates plots like line charts, bar graphs, histograms, and scatter plots. Example: `plt.plot()` for line graphs.