Final Sem questions

1. Introduction to Data Science

a) Definition

Data Science is the field that combines **statistics, computer science, and domain knowledge** to extract **useful information, patterns, and predictions** from data.

b) Key stages in a Data Science project

1. **Problem definition** – What decision or prediction is needed?
2. **Data collection** – From databases, web, sensors, logs, etc.
3. **Data preprocessing** – Cleaning, handling missing values, integration.
4. **Exploratory Data Analysis (EDA)** – Graphs, summaries to understand data.
5. **Model building** – Use ML/statistical models to learn patterns.
6. **Model evaluation** – Test accuracy, precision, error, etc.
7. **Deployment** – Put model into real use (app, dashboard, API).
8. **Monitoring & maintenance** – Update model when data changes.

c) Benefits / advantages

- Better **decision making** (e.g., who to give loan to).
- **Prediction** (sales forecast, churn prediction).
- **Automation** (spam filter, recommendation engines).
- Detecting **fraud and anomalies**.
- **Personalization** (recommend movies, products, content).

---

2. Big Data and Data Science

a) What is Big Data?

Big Data = **datasets so large / fast / complex** that traditional tools (like simple RDBMS) cannot store, process or analyze them efficiently.

b) Characteristics of Big Data (5Vs)

1. **Volume** – Huge size (TB, PB of data).
2. **Velocity** – High speed of generation (streaming, real time).
3. **Variety** – Different types: structured, semi-structured, unstructured.
4. **Veracity** – Uncertainty, noise, inconsistencies in data.
5. **Value** – The actual business benefit extracted from data.

c) Big Data vs Data Science

Aspect	Big Data	Data Science
Focus	Handling large & complex data	Extracting knowledge & predictions from data
Main concern	Storage, processing, scalability	Analysis, modelling, insights
Tools	Hadoop, Spark, NoSQL	Python/R, ML libraries, statistics, BI tools
Relationship	Provides raw material (data)	Uses that data to create value

3. Introduction to Big Data Platforms

A **Big Data platform** is an integrated environment providing **storage, processing, and tools** for massive data.

Key components / technologies

1. **Hadoop Ecosystem**
    - **HDFS (Hadoop Distributed File System)** – Distributed file system to store huge data across clusters.
    - **MapReduce** – Programming model for parallel processing.
    - **YARN** – Resource manager.
    - Tools on top:
        - **Hive** – SQL-like querying.
        - **Pig** – Data flow scripts.
        - **HBase** – NoSQL database.

2. **Apache Spark**
    - In-memory processing (much faster than MapReduce).
    - Supports batch, streaming, ML (MLlib), graph processing.
3. **NoSQL Databases**
    - Handle non-relational, flexible schema data.
    - Examples: MongoDB, Cassandra, CouchDB.

These platforms solve **scalability, speed, and variety** challenges of Big Data.

---

4. Challenges of Conventional Systems

Traditional systems (RDBMS + single server) face issues:

1. **Limited scalability**
    - Hard to store TBs/PBs on single machine.
2. **Performance issues**
    - Queries become very slow on huge tables.
3. **Rigid schema**
    - Data must fit rows & columns; not flexible for JSON, text, images.
4. **Not suited for unstructured data**
    - Emails, PDFs, logs, multimedia are difficult to handle.
5. **High cost**
    - Scaling vertically (bigger machine) is expensive.
6. **Limited fault tolerance**
    - If main server fails → downtime, data loss risk.

Big Data platforms (Hadoop/Spark) address these with **distributed storage & processing**.

---

5. Nature of Data

a) By structure

- **Structured data**

Organized in rows & columns.

Example: SQL tables – Employee(id, name, salary).

- **Semi-structured data**

  Has some structure but not fixed schema.

  Example: JSON, XML, logs.

- **Unstructured data**

  No predefined data model.

  Example: free text, images, audio, video documents.

b) By type

- **Numerical**
  - **Discrete** (countable): number of customers.
  - **Continuous**: height, weight, temperature.
- **Categorical**
  - **Nominal**: labels without order (red/blue, male/female).
  - **Ordinal**: has order (low/medium/high, grade A/B/C).

Understanding nature of data decides **which methods, plots, and models** to use.

---

6. Analytic Processes and Tools

a) Generic Analytics Process

1. **Business understanding** – What is the goal?
2. **Data understanding** – Explore available data.
3. **Data preparation (preprocessing)** – Clean, transform, integrate.
4. **Modelling** – Choose algorithm(s) and build models.
5. **Evaluation** – Compare models, check metrics.
6. **Deployment** – Push model into production (app, API, dashboard).
7. **Feedback/monitoring** – Track performance over time.

b) Tools used in Analytics

- **Programming languages**
  - Python (Pandas, NumPy, Scikit-Learn, Matplotlib).
  - R (tidyverse, ggplot2).
- **Data storage**
  - SQL databases (MySQL, PostgreSQL).
  - NoSQL (MongoDB, Cassandra).
- **Big Data**
  - Hadoop, Spark.
- **Visualization / BI**
  - Power BI, Tableau, QlikView.
- **Others**
  - Excel, Jupyter Notebook.

---

7. Analysis vs Reporting

Reporting

- Focus: **What happened?**
- Uses: **Tables, charts, static dashboards**.
- Example: Monthly sales report, daily traffic report.
- Typically **descriptive** and historical.

Analysis

- Focus: **Why did it happen? What will happen?**
- Uses: statistical tests, ML models, exploratory methods.
- Example: Predict next month's sales, find reasons for churn.
- Helps in **decision making**, optimization, and strategy.

In short: Reporting = telling the story of the past.

Analysis = understanding & predicting the story.

---

8. Modern Data Analytic Tools

a) Business Intelligence (BI) Tools

- **Power BI**
- **Tableau**
- QlikView, Looker, etc.

They allow:

- Easy **data connection** (Excel, SQL, cloud).
- **Drag-and-drop** visual creation.
- Interactive dashboards and filters.

b) Big Data / Processing Tools

- Hadoop, Spark, Kafka.

c) Machine Learning / AI Tools

- Scikit-Learn, TensorFlow, PyTorch, Keras.

---

9. Overview of Power BI

Power BI is a Microsoft BI tool used for **interactive data visualization and reporting**.

Features

- Connects to many sources: Excel, SQL Server, cloud services.
- **Power Query** for data cleaning and transformation.
- **DAX (Data Analysis Expressions)** for calculated columns and measures.
- Build **dashboards** with charts, maps, KPIs, slicers.
- Publish reports to **Power BI Service** for sharing.

Use cases

- Sales dashboards
- Financial reporting

- Marketing campaign analysis

10. Overview of Tableau

Tableau is a powerful visualization tool used for **visual analytics and storytelling**.

Features

- Drag-and-drop interface.
- Works with many data sources.
- **Sheets → Dashboards → Stories** (for presentations).
- Strong support for **maps, advanced charts, filters**.

Tableau vs Power BI (short idea)

- Power BI: cheaper, better with Microsoft ecosystem.
- Tableau: very strong visual and interactive capabilities.

11. Multi-Dimensional Data

Used mainly in **data warehousing** and **OLAP (Online Analytical Processing)**.

Key concepts

- **Dimension** – A perspective or category of analysis.

  Examples: Time, Location, Product, Customer.

- **Measure** – Numeric value being analyzed.

  Examples: Sales, Quantity, Profit.

Example Multi-dimensional view:

Sales measured by **Product × Region × Time**.

OLAP operations

- **Slice** – Fix one dimension and see a subcube.

  Example: Sales for year 2024 only.

- **Dice** – Select a range on multiple dimensions.

Example: Sales for (Bihar, UP) and (2023–2024).

- **Roll-up** – Aggregate to higher level.

  Example: City → State → Country.

- **Drill-down** – Go to more detail.

  Example: Country → State → City → Store.

---

12. Exploratory Data Analysis (EDA)

EDA = First step in analysis where we **explore data visually and statistically** to understand patterns, spot errors, and form hypotheses.

Role of EDA in Data Science

- Detect **missing values**, outliers, anomalies.
- Understand **distribution** of variables.
- Check **relationships** between variables.
- Help choose appropriate **models and transformations**.
- Avoid wrong assumptions.

Basic tools of EDA

a) Plots

- **Histogram** – Distribution of a numeric variable.
- **Box plot** – Median, quartiles, and outliers.
- **Scatter plot** – Relationship between two numeric variables.
- **Bar chart** – Comparison among categories.

b) Graphs

- Line graph – Trends over time.
- Heatmap – Matrix-based color visualization.

c) Summary statistics

- **Central tendency** – Mean, median, mode.

- **Dispersion** – Range, variance, standard deviation, IQR.
- **Shape** – Skewness, kurtosis (sometimes).

13. Need for Data Preprocessing

Real-world data is **not clean**:

- Missing values (NaN, blank cells)
- Duplicate records
- Inconsistent formats (e.g., “India”, “IND”, “IN”)
- Noise and outliers
- Different data sources

Why it is important

- Models trained on dirty data give **wrong or unstable results**.
- Quality of data directly affects **accuracy and reliability**.
- Preprocessing transforms raw data into **consistent, usable form** for analysis and modelling.

14. Data Cleaning

Process of **detecting and correcting** (or removing) errors and inconsistencies.

Common tasks

1. **Handling missing values**
    - Delete rows/columns with too many missing values.
    - Impute using **mean/median/mode**.
    - Use advanced methods like regression or k-NN imputation.
2. **Removing duplicates**
    - Identify duplicate rows/IDs and keep only one.
3. **Correcting incorrect values**
    - Example: Negative age, impossible dates.

4. **Handling outliers**
    - Investigate whether outliers are true or erroneous.
    - Cap or remove them if they are due to error.
5. **Standardizing formats**
    - Date formats (dd-mm-yyyy vs yyyy-mm-dd).
    - Categorical values (Male/Female vs M/F).

Good data cleaning → **high-quality data** → better model and decisions.

---

15. Data Integration and Transformation

a) Data Integration

Combining data from **multiple sources** into a single, unified view.

Examples:

- Merging customer table from CRM + transaction table from sales DB.
- Joining web analytics data + purchase logs.

Challenges:

- Different formats, schemas, key fields.
- **Schema integration** – ensure consistent column names and types.
- **Entity resolution** – recognizing the same entity in different datasets.

b) Data Transformation

Converting data into a suitable format or structure for analysis.

Common transformations:

- **Normalization/Standardization** – scaling numeric values.
- **Aggregation** – summarizing (daily → monthly sales).
- **Encoding** – converting categories to numeric:
    - Label encoding, One-hot encoding.
- **Binning** – converting continuous data to categories.
- **Log / sqrt / power transformations** – to reduce skewness.

Integration + Transformation = **consistent, clean, model-ready data.**

16. Data Reduction

Goal: **reduce the size** of data **without losing important information**.

Why needed?

- Speed up algorithms.
- Reduce storage and computation.
- Remove irrelevant or redundant features.

Techniques

1. **Dimensionality Reduction**
    - Reduce number of attributes/features.
    - Example: **PCA (Principal Component Analysis)**.
    - Removes correlated / less informative features.
2. **Feature Selection**
    - Select only relevant features using:
        - Filter methods (correlation, chi-square).
        - Wrapper methods (forward/backward selection).
        - Embedded methods (LASSO, decision trees).
3. **Numerosity Reduction**
    - **Sampling** – Use a subset of data.
    - **Aggregation** – Group data (e.g., hourly → daily).

17. Discretization & Concept Hierarchy Generation

a) Discretization

Converting **continuous attributes** into **discrete/categorical** intervals.

Example:

Age (continuous) →

- 0–12: Child
- 13–19: Teen
- 20–60: Adult
- 60: Senior

Methods:

- **Equal-width binning** – same interval size.
- **Equal-frequency binning** – same number of records in each bin.
- **Supervised discretization** – uses class labels (entropy-based).

Useful for:

- Decision trees, association rules.
- Simplification and interpretability.

b) Concept Hierarchy Generation

Creating **levels of abstraction** for attributes.

Example 1:

City → State → Country → Continent

Example 2 (Date):

Day → Month → Quarter → Year

Why useful?

- For **roll-up/drill-down** in OLAP.
- To perform analysis at different granular levels.

18. Data Summarization

Summarization gives **compact descriptions** of data.

Methods

1. **Descriptive statistics**
    - Mean, median, mode, variance, standard deviation.
2. **Frequency tables**

- Count of each category.

3. **Grouped summaries**
   - e.g., Average salary per department.
4. **Pivot tables / cross-tabulation**
   - Summarize measures across multiple dimensions.

Purpose:

- Quickly understand main characteristics of data.
- Identify trends and anomalies without going record-by-record.

---

19. Data Normalization

Scaling attribute values to a **common range** or distribution.

Why normalize?

- Many ML algorithms (k-NN, k-means, gradient descent) work better when features are on **similar scale**.
- Avoid dominance of features with large numerical range.

Common methods (formulas ok, but no numeric problems)

1. **Min–Max Normalization**

Rescales data to [0, 1] (or any [a, b]).

x′=x−xminxmax−xminx' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}

x′=xmax−xminx−xmin

1. **Z-score (Standardization)**

Centers data around mean with unit variance.

x′=x−μσx' = \frac{x - \mu}{\sigma}

x′=σx−μ

1. **Decimal Scaling**

Move decimal point to bring values into a smaller range.

x′=x10jx' = \frac{x}{10^j}

x′=10jx

20. Explain Computer Vision and its real-world applications

Computer Vision is a field of Artificial Intelligence that enables machines to *see, interpret, and make decisions* based on visual data such as images and videos.

How it works

- Image acquisition
- Preprocessing (noise removal, resizing)
- Feature extraction
- Classification / detection using ML or deep learning

Real-world applications

1. **Facial Recognition** – Phone unlock, airport security
2. **Autonomous Vehicles** – Lane detection, traffic sign recognition
3. **Medical Imaging** – Detecting tumors in X-rays or MRIs
4. **Surveillance Systems** – Tracking suspicious activity
5. **Industrial Automation** – Detecting defects in manufacturing
6. **Retail Analytics** – Counting customers, analyzing behavior
7. **Agriculture** – Crop disease detection

21. Concept of Population and Sample in Statistical Inference

Population

The *entire group* of items, people, or events you want to study.

Example: All students in India.

Sample

A *subset* of the population used for analysis.

Example: 500 students selected from Indian colleges.

Why samples are used

- Studying entire population is expensive and time-consuming
- Sampling allows statistical inference with controlled error

Statistical inference

Using sample information to draw conclusions (estimates, predictions) about the population.

22. Explain Streaming Data with example

Streaming Data is continuous, real-time data generated at high speed.

Characteristics

- Continuous flow
- Requires fast processing
- Often unbounded and time-sensitive

Examples

- **Stock market price updates** (streaming every second)
- **Sensor data** from IoT devices
- **Live user clicks** on websites
- **Social media feeds** (tweets, likes)
- **Real-time GPS tracking** in ride-sharing apps

Tools used: Apache Kafka, Spark Streaming, Flink.

23. Explain Feature Selection and its methods

Feature Selection is the process of selecting the most relevant attributes from the dataset to improve model performance.

Benefits

- Reduces overfitting
- Improves model accuracy
- Reduces training time
- Simplifies models

Methods of Feature Selection

A. Filter Methods

Select features based on statistical scores.

Examples:

- Correlation coefficient
- Chi-square test
- ANOVA
- Mutual information

B. Wrapper Methods

Use machine-learning models to evaluate feature subsets.

Examples:

- Forward selection
- Backward elimination
- Recursive Feature Elimination (RFE)

C. Embedded Methods

Feature selection happens during model training.

Examples:

- LASSO (L1 Regularization)
- Decision tree feature importance

---

24. What is Model Fitting in Statistics or ML?

Model fitting means **training a model** so that it learns the pattern from given data.

Types of fitting

- **Underfitting**: Model is too simple → poor accuracy
- **Overfitting**: Model memorizes training data → poor generalization
- **Good fit**: Model captures real patterns and performs well on unseen data

Goal

Achieve a balance between bias and variance for accurate predictions.

25. Explain Time Series Analysis

Time Series Analysis (TSA) studies data collected over time intervals.

Characteristics

- Time-dependent
- Shows trends, seasonality, cycles

Components of Time Series

- **Trend**: Long-term increase/decrease
- **Seasonality**: Repeating patterns (e.g., festival sales)
- **Cyclic variations**: Business cycles
- **Irregular variations**: Random fluctuations

Applications

- Sales forecasting
- Weather prediction
- Stock price analysis
- Economic planning

26. Market Basket Analysis in Association Rule Mining

Market Basket Analysis (MBA) identifies items frequently bought together.

Example: Customers buying **bread** often buy **butter**.

It uses:

- **Frequent itemsets**
- **Association rules** with **support, confidence, lift**

Applications

- Cross-selling in retail
- Recommendation systems
- Product placement in stores

27. Process of Knowledge Discovery in Database (KDD)

KDD is the full process of extracting useful knowledge from data.

Steps

1. **Data Cleaning** – Remove noise, missing values
2. **Data Integration** – Combine multiple sources
3. **Data Selection** – Choose relevant data
4. **Data Transformation** – Normalize, aggregate
5. **Data Mining** – Apply algorithms (classification, clustering, ARM)
6. **Pattern Evaluation** – Identify useful patterns
7. **Knowledge Presentation** – Visualization and interpretation

28. Steps to Generate Frequent Itemsets using Apriori Algorithm

1. **Generate C1** – List all candidate 1-itemsets
2. **Compute L1** – Keep only frequent ones (support ≥ min support)
3. **Generate C2** – Join L1 with L1 to create 2-item candidates
4. **Scan database** – Count support for C2
5. **Generate L2** – Keep only frequent 2-itemsets
6. **Repeat** joining Lk to create Ck+1
7. **Stop** when no new frequent itemsets can be generated

Uses "**Apriori property**":

If an itemset is frequent, all its subsets must be frequent.

---

29. Why FP-Growth is more efficient than Apriori?

1. **No candidate generation**

   Apriori generates many candidates; FP-growth eliminates this.

2. **Compresses data into FP-tree**

   Reduces repeated scanning.

3. **Requires only 2 database scans**

   Apriori requires multiple scans → slower.

4. **Memory-efficient**

   FP-tree stores frequency patterns compactly.

5. **Faster for large datasets**

   Excellent for high-dimensional data.

---

30. What is an Outlier? Explain Outlier Analysis Methods

Outlier

A data point that is *significantly different* from other observations.

Causes

- Measurement error
- Fraud / rare events
- Noise

Outlier Detection Methods

1. Statistical Methods

- Z-score
- IQR (Inter-Quartile Range) method
- Boxplot analysis

2. Distance-based Methods

- k-NN distance
- DBSCAN clustering

3. Density-based Methods

- LOF (Local Outlier Factor)

4. Model-based Methods

- Isolation Forest
- Autoencoders

---

31. Main Steps of FP-Growth Algorithm

1. **Scan database once** to build:
    - Frequent item list
    - FP-tree (compressed representation)
2. **Construct conditional pattern base** for each item

3. **Build conditional FP-tree**
4. **Generate frequent itemsets** by recursively exploring FP-trees

32. Define: Decision Tree

A **Decision Tree** is a supervised learning model that makes decisions by splitting data into branches based on feature values.

Key components

- **Root node** – First decision point
- **Internal nodes** – Feature-based splits
- **Leaf nodes –** Final predicted class/value

Used for: classification, regression, rule extraction.

33. Explain types of Tokenization and their application

Tokenization = Splitting text into meaningful units (tokens).

Types

1. Word Tokenization

Splits sentence into individual words.

Application: Sentiment analysis, Information retrieval.

2. Sentence Tokenization

Splits paragraph into sentences.

Application: Text summarization.

3. Character Tokenization

Splits text into characters.

Application: Language modeling for small datasets.

4. Subword Tokenization (BPE, WordPiece)

Breaks uncommon words into smaller pieces.

Application: Transformers (BERT, GPT).

34. Define: Sentiment Analysis

Sentiment Analysis identifies the **emotion or opinion** expressed in text.

Types:

- **Positive**
- **Negative**
- **Neutral**

Applications:

- Customer review analysis
- Social media monitoring
- Brand perception tracking

35. What is an RNN? How is it different from Feedforward NN?

Recurrent Neural Network (RNN)

A neural network designed for sequential data. It has **feedback connections** that allow information to persist.

Difference from Feedforward NN

Feedforward NN	RNN
No memory of previous inputs	Remembers previous inputs using hidden state
Suitable for independent data	Suitable for sequences (text, speech, time series)
Processes input once	Processes input recursively

Applications: language modelling, speech recognition.

36. What is the Vanishing Gradient Problem in RNNs?

During training with backpropagation through time (BPTT), gradients become **extremely small**, causing:

- Slow learning
- Inability to capture long-term dependencies
- Model forgets early information

This is why basic RNNs struggle with long sequences.

37. Define: NLP

Natural Language Processing (NLP) is the field of AI that enables computers to understand, process, and generate human language.

Applications:

- Chatbots
- Translation
- Text classification
- Speech recognition

38. Explain Text Lemmatization

Lemmatization reduces words to their **base or dictionary form (lemma)**.

Examples:

- “Running”, “ran”, “runs” → “run”
- “Better” → “good”

Lemmatization uses grammar and vocabulary → more accurate than stemming.

39. What are the Input, Forget, and Output Gates in LSTM?

LSTM contains three gates:

1. Input Gate

Controls how much new information enters the memory cell.

2. Forget Gate

Decides what information to remove from memory.

3. Output Gate

Controls what information is sent out as the next hidden state.

Each gate uses a sigmoid function to allow/select information.

40. How does LSTM address the vanishing gradient problem?

- LSTM uses **cell state**, which allows gradients to flow unchanged.
- Gates control information flow, preventing gradients from shrinking.
- Memory cell preserves long-term dependencies.

Therefore, LSTMs can learn long sequences better than RNNs.

41. Difference between Classification and Clustering

Classification (Supervised)	Clustering (Unsupervised)
Uses labeled data	No labels
Predicts class/category	Groups similar items
Examples: spam detection, disease prediction	Customer segmentation

42. Define Machine Learning and explain its types

Machine Learning

A field of AI that trains machines to learn patterns from data without explicit programming.

Types of ML

1. Supervised Learning

Uses labeled data.

Examples: Classification, Regression.

2. Unsupervised Learning

No labels.

Examples: Clustering, Dimensionality reduction.

3. Reinforcement Learning

Agent learns by reward/punishment.

Examples: game playing, robotics.

---

43. Define: Neural Network, K-Means, Naïve Bayes

A. Neural Network

A computational model inspired by the human brain. Consists of layers of neurons that learn patterns from data.

B. K-Means Algorithm

Unsupervised clustering algorithm that divides data into **K clusters** by minimizing the distance between points and cluster centers.

C. Naïve Bayes

A probabilistic classifier based on Bayes' Theorem assuming **features are independent**.

Used for: spam detection, text classification.