DATA ANALYTICS

BRANCH:- OPEN ELECTIVE

SEMESTER – FIFTH

These important questions have been prepared using your previous exam papers (PYQs), verified concepts, and additional reference from trusted online academic sources.
For deeper understanding, please refer to your class notes as well.

1️⃣ HIGH & LONG IMPORTANT QUESTIONS

1. **Define Artificial Intelligence (AI).** Explain the different **types of AI** (Narrow, General, Super AI) and discuss the major **applications of AI** in the real world (Healthcare, Finance, E-commerce).
2. **What is Machine Learning (ML)?** Explain the **Machine Learning Workflow/Pipeline** (Data Collection → Preprocessing → Model Training → Evaluation → Deployment) with a neat block diagram.

3. **Difference between Supervised, Unsupervised, and Reinforcement Learning.** Give examples for each.
4. **What is Data Preprocessing?** Why is it important? Explain the steps to handle **Missing Values** and **Outliers** in a dataset.
5. **Explain Linear Regression.** What are its assumptions? Differentiate between Simple Linear Regression and Multiple Linear Regression.
6. **What is a Decision Tree?** Explain key terms like **Root Node, Leaf Node, Entropy, and Information Gain**. How do we prevent overfitting in Decision Trees (Pruning)?
7. **Define Deep Learning.** Explain the architecture of an **Artificial Neural Network (ANN)** including Input Layer, Hidden Layer, and Output Layer. How is it different from biological neurons?

8. **What is Cloud Computing in AI?** Explain the difference between **IaaS, PaaS, and SaaS** service models.
9. **Explain the Confusion Matrix** and how to calculate the following metrics:
    - Accuracy
    - Precision
    - Recall
    - F1-Score
10. **What is Natural Language Processing (NLP)?** Explain the steps of text processing: **Tokenization, Stemming, Lemmatization, and Stop Word Removal**.
11. **Explain the K-Means Clustering algorithm.** How do we decide the number of clusters (K)?

12. **What is Docker?** Explain the architecture of Docker (Client, Daemon, Images, Containers) and why containers are used in AI deployment.

2️ IMPORTANT & SHORT QUESTIONS

1. **Differentiate between:**
   - AI vs. Machine Learning vs. Deep Learning.

2. **Python Libraries:** Briefly explain the use of the following libraries in AI:
   - **NumPy** (for numerical data)
   - **Pandas** (for data manipulation)
   - **Matplotlib/Seaborn** (for visualization)
   - **Scikit-learn** (for ML models)
3. **What is Big Data?** Explain the **5 Vs of Big Data** (Volume, Velocity, Variety, Veracity, Value).
4. **Activation Functions:** Define and draw the graph for:
   - Sigmoid
   - ReLU (Rectified Linear Unit)
   - Tanh
5. **What is Version Control System (VCS)?** Why is **Git** and **GitHub** useful for AI projects?
6. **Data Visualization:** Explain the difference between a **Histogram** and a **Scatter Plot**. Why do we visualize data?
7. **What is a Hypothesis?** Explain Null Hypothesis and P-value in short.
8. **Define Sentiment Analysis.** Give one real-world business example of it.
9. **What is MLOps?** Why is it needed in the industry?
10. **Explain SVM (Support Vector Machine).** What is a Hyperplane and Support Vector?

3️ "AA BHI SAKTA HAI" QUESTIONS (20–30% Probability)

(Tricky or new syllabus topics. Read them once if you want to score Top grades.)

1. **Ethics in AI:** Discuss the ethical challenges in AI like **Bias, Privacy, and Job displacement**.
2. **Ensemble Learning:** Write a short note on **Bagging (Random Forest)** and **Boosting**. Why are they better than single models?
3. **Bayes' Theorem:** State the theorem and its application in Naive Bayes Classification.
4. **Dimensionality Reduction:** What is **PCA (Principal Component Analysis)** and why is it used to reduce data dimensions?
5. **TF-IDF:** What does Term Frequency-Inverse Document Frequency mean in NLP?
6. **Maths for AI:** Briefly explain the concept of **Eigenvalues and Eigenvectors** (Linear Algebra) and their use in AI.
7. **Data Integration:** What are the challenges in integrating data from different sources?

8. **Backpropagation:** Briefly explain how a Neural Network "learns" by updating weights to reduce loss.

💡 ***Analyst Tip for JUT Exam:***

Since this subject (AI&ML) has a lot of "practical" topics like Python and Docker in the syllabus, **do not write code** in the exam unless specifically asked (which is rare in theory exams). Instead, **describe the concept**.

- *Example:* Instead of writing code to clean data, write: *"Data cleaning involves removing null values using methods like mean imputation..."*
- **Diagrams are key:** Draw the Neural Network layers, the ML pipeline blocks, and the Confusion Matrix table. This guarantees good marks in JUT.

QUICK REVISE

⚡ **1. AI Fundamentals & Setup (Weeks 1, 2)**

Topic	Quick Note
What is AI?	Making computers smart like humans to solve problems, learn, and decide.
Types of AI	**Weak/Narrow AI:** Does *one* specific task (e.g., Google Search, Alexa). **Strong/General AI:** Can do *any* intellectual task (future goal).
AI SDLC	The process is **data-centric** and involves continuous **monitoring** and **retraining** of the model, unlike traditional software development.
Ethics in AI	**Fairness** (avoiding bias), **Privacy** (protecting data), and **Accountability** (fixing errors).
Version Control (Git)	A system to track and manage changes in your code/data over time, allowing easy rollback and teamwork.

☁️ **2. ML, Cloud & Data Basics (Week 3, 5)**

Topic	Quick Note
Machine Learning	Teaching computers to learn patterns from data **without explicit programming**.
ML Types	**Supervised:** Learning from labeled examples (Input + Correct Answer). **Unsupervised:** Finding hidden groups/patterns in unlabeled data.
ML Pipeline	Data Collection > Data Preprocessing > Model Training > Model Testing > Deployment.
Cloud Models	**IaaS:** Renting the **hardware** (servers, storage). **PaaS:** Renting the **platform** (OS, tools to build). **SaaS:** Renting the **ready-made software** (Gmail, CRM).
Big Data (5 Vs)	**Volume** (Large size), **Velocity** (High speed), **Variety** (Different formats), **Veracity** (Accuracy), **Value** (Usefulness).
NumPy & Pandas	**NumPy:** Used for fast calculation on arrays and matrices. **Pandas:** Used for manipulating table-like data (DataFrames).

📊 3. Data Exploration & Math (Week 4, 6)

Topic	Quick Note
Data Visualization	Presenting data visually (graphs/charts) to easily find **patterns, trends, and outliers**.
Central Limit Theorem	Even if the original data is weird, the average of many samples tends to form a **Normal Distribution** (Bell Curve).
Bayes' Theorem	Formula to calculate the **probability of an event** based on information we already know (prior knowledge).
Correlation	Measures how strongly two variables are **related** (from -1 to +1).
Vectors & Matrices	Data and parameters in AI are stored as **Vectors** (1D array) and **Matrices** (2D array) for efficient processing.
Eigenvalues/vectors	Special numbers/vectors that show the **direction of maximum variance** (used in PCA for dimension reduction).

🩹 4. Data Preprocessing (Week 7)

Topic	Quick Note
Missing Values	Data entries that are empty. Handled by **imputation** (filling with mean/median) or **deletion**.
Outliers	Data points that are far away from the rest of the data. Handled by **removal** or using **log transformation**.
Normalization	Scaling data values to a small, fixed range (usually **0 to 1**).
Standardization	Scaling data so that the mean is **0** and the standard deviation is **1** (Z-Score).
Data Reduction	Making the data smaller (reducing rows or features) while keeping important information.

📉 5. Regression & Evaluation (Week 8)

Topic	Quick Note
Data Splitting	Dividing data into **Training** (for learning), **Validation** (for tuning), and **Testing** (for final grade).
Overfitting	Model learns the **training data too perfectly** (memorizes noise), performs badly on new data.
Underfitting	Model is **too simple**, cannot capture the basic patterns of the data.
Linear Regression	Used to **predict a continuous value** (like price, temperature) based on input features using a straight line equation.
R-Squared	Measures how well the regression line **fits the observed data** (higher is better, close to 1).
RMSE	Measures the **average size of the errors** (difference between predicted and actual value).
Cross-Validation	Method (like K-Fold) to repeatedly train and test a model on different subsets of data to get a **reliable performance estimate**.

6. Classification & Ensemble (Week 9)

Topic	Quick Note
Classification	Used to **predict a category/class** (like Yes/No, Spam/Not Spam).
Decision Tree	A flow-chart model that makes decisions by asking a series of **Yes/No questions** on features.
Entropy	Measures the **impurity** or randomness in a group of data.
Information Gain	The drop in entropy after a split; used to decide the **best feature** to split the tree on.
Logistic Regression	Used for **Binary Classification** (Yes/No), uses the Sigmoid function to output probability (0 to 1).
Confusion Matrix	A table summarizing the model's predictions: **TP, TN, FP, FN**.

🤖 7. Deep Learning, Unsupervised & MLOps (Week 10)

Topic	Quick Note
Deep Learning	Uses **Neural Networks** with many hidden layers to solve complex problems, especially with unstructured data (images, text).
Perceptron	The basic unit of a Neural Network. It takes weighted inputs, sums them up, and passes them through an **Activation Function**.
Activation Function	Introduces **non-linearity** (e.g., ReLU, Sigmoid) so the network can learn complex, non-straight line relationships.
Backpropagation	The main learning algorithm. It calculates the **error** and sends it backward to **adjust the weights** of the network layers.
K-Means Clustering	Unsupervised algorithm that groups data points into **K number of clusters** based on how close they are to the cluster center.
Dimensionality Reduction (PCA)	Reduces the number of features/columns by finding the most important ones, helping to **speed up training**.

Topic	Quick Note
MLOps	The set of practices for **deploying, monitoring, and managing** ML models in a real-world production environment.

💬 8. NLP & Deployment (Week 11, 12)

Topic	Quick Note
NLP	Giving computers the ability to **understand and process human language** (text and speech).
Tokenization	Breaking down text into small meaningful units (words or phrases).
Stemming	Chopping off the end of words to get a rough root word (e.g., *running* > *run*).
Lemmatization	Getting the proper dictionary base word (lemma) using vocabulary rules (e.g., *better* > *good*).
TF-IDF	A score that tells you how **important a word** is in a specific document compared to all other documents.
Sentiment Analysis	Determining the **mood or opinion** (positive, negative, neutral) of a piece of text.
Docker	A tool to package an application and all its dependencies into a single, reliable unit called a **container**.
Container	Ensures that the trained ML model runs **exactly the same way** on a developer's computer, a tester's computer, or the final server.