UNIT 1

1. Explain the DIKW Pyramid (Data, Information, Knowledge, Wisdom) with suitable examples.

The DIKW Pyramid is a hierarchy representing the relationship between data, information, knowledge, and wisdom.

- Data: Raw facts or figures without context. Example: "120, 145, 160"
- Information: Processed data with meaning. Example: "The average blood pressure readings this week are 120, 145, and 160."
- Knowledge: Interpretation of information. Example: "Blood pressure readings over 140 are considered high."
- Wisdom: Application of knowledge to make informed decisions. Example: "The patient should be advised to consult a doctor for hypertension treatment."

2. Describe the different stages in the Data Lifecycle and their significance in Data Science.

The data lifecycle includes:

1. Data Generation: Data is produced through devices, users, transactions, etc.
2. Data Collection: Gathering data from multiple sources.
3. Data Storage: Storing in databases or data lakes.
4. Data Processing: Cleaning, transforming data into usable form.
5. Data Analysis: Extracting insights using statistical or ML methods.
6. Data Visualization: Representing insights through graphs and dashboards.
7. Data Interpretation and Decision Making.
8. Data Archival/Destruction: Secure storage or deletion.

Each stage ensures data is managed effectively for decision-making.

3. Discuss the key roles in a Data Science project.

- Data Scientists: Analyze data, build models, derive insights.

- Data Engineers: Build pipelines, handle data storage and processing.

- ML Engineers: Optimize and deploy machine learning models.

All collaborate to ensure data flows from raw to insights effectively.

4. Ethical considerations in Data Science?

- Privacy: Ensure personal data is not misused. Ex: GDPR compliance.

- Bias: Biased data leads to unfair models. Ex: Hiring tools preferring one gender.

- Fairness: Treat all individuals equitably.

- Transparency and Accountability: Explainable models and decisions are critical.

5. Evolution of Data Science.

Traditional statistics focused on data collection and hypothesis testing.

With increased computing power:

- Business Intelligence emerged

- Machine Learning models automated pattern detection

- AI introduced deep learning and NLP

Industries like healthcare, finance, marketing now rely on predictive analytics and AI-driven automation.

UNIT 2

1. Define structured, semi-structured, and unstructured data with suitable examples.

- Structured Data: Organized in tables with rows and columns. Example: SQL databases.

- Semi-structured Data: Doesnt conform to relational databases but has tags or markers. Example: JSON, XML.

- Unstructured Data: No predefined format. Example: Images, videos, social media posts.

2. Explain any two data collection methods in detail.

1. Surveys:

   - Advantages: Cost-effective, scalable

- Disadvantages: May suffer from response bias

2. Sensors:

   - Advantages: Real-time, automated data

   - Disadvantages: Expensive setup, limited to certain domains

3. Differences between primary and secondary data sources.

- Primary Data: Collected first-hand. Example: Experiment results, surveys.

- Secondary Data: Previously collected for other purposes. Example: Government databases, research articles.

4. Web scraping for data collection.

Web scraping involves using bots to extract data from websites.

Uses: Price comparison, sentiment analysis.

Ethical concerns:

- Terms of Service violations

- User data privacy

- Legal regulations (e.g., GDPR)

5. Importance of data pre-processing.

Pre-processing prepares data for analysis. Essential to improve model accuracy.

Techniques:

1. Normalization: Scaling features.

2. Handling missing data: Imputation or removal of null values.

UNIT 3

1. Define model representation in data science.

Model representation refers to the formal mathematical structure used to map input data to outputs.

Example: A linear regression model represents data with $y = mx + b$.

2. Differentiate between statistical, ML, and deep learning models.

- Statistical: Based on probability theory. Ex: Linear regression

- ML: Learn patterns from data. Ex: Decision trees

- DL: Neural networks for complex data. Ex: CNN for images

3. Training, validation, and test data.

- Training: Used to fit the model

- Validation: Fine-tune hyperparameters

- Test: Evaluate model performance on unseen data

4. Overfitting vs Underfitting.

- Overfitting: Model too complex, fits noise.

- Underfitting: Model too simple, misses patterns.

Detection: Learning curves

Mitigation: Cross-validation, regularization

5. Bias-variance tradeoff.

High bias: Underfitting, simplistic models.

High variance: Overfitting, sensitive to training data.

Goal: Find balance for best generalization.

UNIT 4

1. Why is model evaluation important?

Evaluation determines how well a model generalizes to new data.

Example: Classifier with 95% accuracy may still misclassify minority classes.

2. Different evaluation metrics per ML problem.

- Classification: Accuracy, Precision, Recall

- Regression: MAE, RMSE

- Clustering: Silhouette Score, Dunn Index

3. Accuracy, Precision, Recall, F1 Score.

- Accuracy: Correct predictions / Total

- Precision: TP / (TP + FP)

- Recall: TP / (TP + FN)

- F1 Score: Harmonic mean of precision and recall

4. MAE, MSE, RMSE, R2 in regression.

- MAE: Mean Absolute Error, simple

- MSE: Mean Squared Error, penalizes large errors

- RMSE: Root of MSE, interpretable

- R2: Proportion of variance explained

5. Evaluation for imbalanced datasets.

- Use Precision-Recall curve and F1 Score over accuracy

- Helps highlight minority class performance

- Example: Fraud detection, medical diagnosis

yashsoni