Steps a Data Scientist Follows When Given a Dataset

1. Understand the Problem

Meet with stakeholders to understand the business problem or research question. Define objectives and success metrics (e.g., classification accuracy, RMSE).

2. Acquire the Data

Load data from the given dataset (CSV, Excel, database, etc.). Merge with additional data sources if necessary.

3. Data Cleaning & Preprocessing

Handle missing values, fix data types, remove duplicates/outliers, normalize or scale values, and encode categorical variables.

4. Exploratory Data Analysis (EDA)

Use visualizations and statistics to understand patterns, trends, relationships, and anomalies.

5. Feature Engineering

Create new features, drop irrelevant ones, and apply dimensionality reduction if necessary.

6. Model Selection

Choose appropriate ML algorithms and split data into training/testing sets. Use cross-validation if needed.

7. Model Training

Fit the chosen model(s) on the training data and tune hyperparameters.

8. Model Evaluation

Evaluate models using metrics like accuracy, precision, recall, F1-score, RMSE, etc., and visualize results.

9. Model Improvement

Try different algorithms, ensembles, and revisit feature engineering and hyperparameters.

10. Reporting & Interpretation

Create visualizations and reports for stakeholders, interpret results in context, and use dashboards if needed.

11. Deployment (Optional)

Deploy models using APIs or cloud platforms, and monitor performance.

12. Feedback Loop

Collect new data or feedback, and update/retrain the model as needed.