IDA-IMP

Module 1: Introductory Concepts

1. **Overview of Data Science and Data Analytics:** Data Science is described as a multidisciplinary field that uses methods to analyze data and extract useful knowledge. It combines multiple disciplines such as Mathematics, Statistics, Computer Science, and Machine Learning. Data analytics refers to the process and practice of analyzing data to answer questions, extract insights, and identify trends using various tools and techniques.
2. **Types of Analytics:** There are four major types of analytics.
   - **Descriptive analytics** examines data to understand what has already happened. It identifies trends and relationships from raw data and describes what happened or is happening. It answers questions like "What happened?".
   - **Diagnostic analytics** goes deeper to understand *why* something happened. This involves comparing trends, uncovering correlations, and determining causal relationships.
   - **Predictive analytics** uses historical data, past trends, and assumptions to answer questions about what *will* happen in the future.
   - **Prescriptive analytics** aims to identify specific actions that should be taken to reach future goals.
3. **Data Science Process:** A data science process involves several stages and activities to find a solution for a business problem. The process steps include: Framing the Business Problem, Data Acquisition, Data Preparation, Data Exploration (EDA), Feature Engineering, Model Building, and Presentation and Model Deployment. Each step involves specific tasks, such as asking the right questions to define the problem, collecting required data, cleaning and enhancing data quality, understanding data through statistics and visualization, transforming data and selecting features, training and evaluating models, and presenting/deploying the results.

Module 2: Data Preparation

1. **Importance of Data Pre-processing:** Data pre-processing is crucial because real-world data is often incomplete, noisy, and inconsistent. Poor data quality leads to inaccurate insights and unreliable results; models built on flawed data will produce incorrect predictions. This phase enhances data quality and prepares it for subsequent analysis and modeling steps. Key data quality factors include Accuracy, Consistency, Completeness, Validity, Timeliness, and Uniqueness.
2. **Major Tasks: Data Cleaning, Integration, and Transformation:**
   - **Data Cleaning** involves handling missing values, removing outliers, correcting inconsistent data, and smoothing noisy data to provide complete and accurate samples.
   - **Data Integration** is the process of combining data from different sources, which may be heterogeneous in nature, to create a coherent and unified data store. Issues include schema integration, entity identification, redundancy, data value conflicts, and tuple duplication.
   - **Data Transformation** converts data from one format to another, changing its structure, format, or values to make it suitable for analysis and modeling. Methods include data aggregation, discretization (converting continuous data to discrete bins), normalization (scaling numerical features), and generalization (converting low-level data to high-level concepts).
3. **Dimensionality Reduction:** This involves reducing the number of attributes or features in a dataset. It is necessary due to the "curse of dimensionality," where high-dimensional data can lead to complex models, make visualization difficult, and increase computational cost. Correlated or redundant variables can also affect model training. Techniques include removing attributes with high missing value ratios or low variance, filtering based on high correlation, and using methods like Principal Component Analysis (PCA), Backward Feature Elimination, or Forward Feature Selection. PCA is a technique that extracts a smaller set of transformed variables (Principal Components) that still contain most of the information from the original large set.

Module 3: Exploratory Data Analysis (EDA)

1. **What is EDA?:** Exploratory data analysis (EDA) is the first step in the data discovery process, involving preliminary analysis and investigation of data sets. Its goal is to summarize main characteristics, often using summary statistics and data visualization. EDA helps gain a better understanding of the data, uncover patterns, determine relationships between variables, identify important variables, and detect outliers and anomalies.
2. **Categories and Techniques:** EDA methods are cross-classified as either non-graphical (calculation of summary statistics) or graphical (visualizing data), and as either univariate (looking at one variable) or multivariate (looking at two or more variables). This results in four categories: univariate non-graphical (e.g., frequency tabulation for categorical data, calculating central tendency and variability for quantitative data), multivariate non-graphical (e.g., cross-tabulation for categorical data, correlation matrices for quantitative data), univariate graphical (e.g., Histograms and Boxplots), and multivariate graphical (e.g., Scattered Plots, Heat Maps, Grouped Bar graphs, Side-by-side Boxplots).
3. **Key Summary Statistics:** Understanding quantitative data involves examining its central tendency and variability. Measures of central tendency include the **Mean** (average value), **Median** (middle value in a sorted list, robust to outliers), and **Mode** (most frequent value). These measures can indicate if the data is symmetrically distributed or skewed. Measures of variability or spread describe how far data values lie from each other and the center. These include **Standard Deviation** (average distance from the mean) and **Interquartile Range (IQR)** (the range of the middle 50% of data, calculated as Q3 - Q1). Skewness (asymmetry) and Kurtosis (peakedness) are also univariate descriptors.

Module 4: Machine Learning and Model Building

1. **Types of Machine Learning:** Machine learning enables computers to learn from data and improve performance without being explicitly programmed. The main types include:
   - **Supervised Learning:** The program is trained on labeled data (input-output pairs) to learn a mapping function (model) that predicts labels for unseen data. Examples include Classification and Regression.
   - **Unsupervised Learning:** The program learns from unlabeled data, finding patterns or groupings without predefined output labels. Examples include Clustering and Association Rule Mining.
   - **Reinforcement Learning:** (Mentioned, but not detailed in the sources).
   - **Semi-supervised Learning:** Used when limited labeled data is available, often labeling unlabeled data before training a supervised model.
2. **Supervised Learning Techniques: Regression and Classification:**

- **Regression** is used for predicting a continuous output variable (response) based on one or more input variables (predictors). Linear Regression assumes a linear relationship between inputs and the output.
- **Classification** is used for predicting a categorical output label (class variable) for unseen data based on training data where each example is labeled with a predefined class. This can be Binary (two classes, e.g., spam or not) or Multi-class (more than two classes).

3. **Unsupervised Learning Techniques: Clustering and Association Rule Mining:**
   - **Clustering** (Cluster Analysis) is the process of grouping similar data objects into subsets called clusters. The goal is to achieve high similarity within clusters (intra-class) and low similarity between clusters (inter-class). The k-Means algorithm is a popular partitioning approach that divides data into k clusters by iteratively assigning points to the nearest centroid and updating centroids.
   - **Association Rule Mining** is a method for discovering relationships (rules) between variables in large databases, typically transaction data. Rules are in the form $X \rightarrow Y$, meaning when itemset X occurs, itemset Y occurs with certain probability. Key concepts include Support (frequency of itemset occurrence) and Confidence (how often Y appears when X is present).

Module 5: Model Evaluation

1. **Assessing Model Accuracy and Quality of Fit:** It's important to quantify how well a model's predictions match observed data. For regression models, a common measure of the quality of fit is the **Mean Squared Error (MSE)**, which is the mean of the squared differences between predicted and observed values. MSE can be calculated for training data (training MSE) or unseen test data (test MSE).
2. **The Bias-Variance Trade-off:** The expected test MSE can be decomposed into Variance, Squared Bias, and Irreducible Error.
   - **Bias** is the error introduced by approximating a complex real-life problem with a simpler model. High bias occurs when the model is too simple and doesn't pay enough attention to the training data, leading to high errors on both training and test data (**Underfitting**).
   - **Variance** is the amount the model's estimate of the function would change if a different training dataset were used. High variance occurs when the model is too complex and pays too much attention to the training data, failing to generalize well to unseen data (**Overfitting**).
   - The goal is to find a balance between bias and variance to minimize the total expected test error.
3. **Model Evaluation Metrics and Methods:** For classification models, a common tool is the **Confusion Matrix**, which summarizes the counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Key metrics derived from the confusion matrix include **Accuracy** ((TP+TN)/(Total)), **Precision** (TP/(TP+FP)), **Recall** (TP/(TP+FN)), and **F1 Score** (a combination of precision and recall). Sensitivity and Specificity are also used, particularly for class imbalance problems. Common methods for evaluating models include **Cross-validation** (e.g., k-fold cross-validation), where the data is partitioned and the model is trained and tested multiple times on different partitions. Visualizations like the **Residual Plot** (for regression) and the **ROC Curve** (for classification) can also be used to assess model performance.