**Introduction for Data Science**

**Week-2**

**Long Descriptive Questions**

**1. Explain about data engineering and List out the steps of data engineering process.**

Data engineering is an important aspect of data management that involves designing, building and maintaining systems and infrastructure to collect, store, process and distribute data to various data users, such as data analysts, data scientists and business users

**Data Ingestion:**

* The process starts with collecting data from various sources, which can be databases, APIs, logs, flat files, and more.
* Data engineers must develop strategies for reliably and efficiently loading data into data warehouse systems.

**Data Storage:**

* Once data is collected, it must be stored in a high-quality format for retrieval and analysis.
* Common storage technologies include relational databases, data warehouses, data lakes, and MySQL databases.

**Data Cleaning and Preprocessing:**

* The raw data generally contain errors, missing values, and inconsistencies.
* Data engineers must prepare and pre-process data to ensure it is quality and ready for analysis. This includes functions such as data validation, transformation and enrichment.

**Data Transformation:**

* Data may need to be combined or transformed with other data types to meet specific business needs.
* Typically, ETL (Extract, Transform, Load) programs are used to process data.

**Data Integration:**

* Often data comes from different sources, and it’s important to integrate and standardize this data to get an integrated view.
* Integration can include connecting data systems, creating master data records, and ensuring data consistency.

**Data Orchestration:**

* Data engineering pipelines must be developed to ensure the smooth flow of data from source to destination.
* Workflow management tools are used to automate and schedule data processes.

**Data Quality Assurance:**

* Continuous monitoring and validation of data quality is essential to ensure that data remains accurate and reliable over time.
* Data engineers establish processes and analytics to identify and resolve data quality issues.

**Data Delivery and Serving:**

* Once the data is processed and ready, it must be delivered to recipients in a usable format.
* Data engineers create data APIs, dashboards, or other interfaces for users to access the data

**2. Brief about data science pipeline**

A data science pipeline is a structured and systematic approach to managing the end-to-end process of extracting insights and knowledge from data. It involves a series of well-defined steps that data scientists and analysts follow

**Data Collection:**

* Identify and collect relevant data from a variety of sources, such as databases, APIs, files, or web scraping.
* Ensure data quality and accuracy, by addressing issues such as missing values ​​and outliers.

**Data Preprocessing**:

* Prepare and preprocess the data to make it suitable for analysis.
* Deal with missing data, outliers, and duplicate records.
* Perform data normalization, scaling, and encoding of categorical variables.

**Exploratory Data Analysis (EDA):**

* Analyze and visualize the data to better understand its characteristics.
* Identify patterns, trends, and relationships in the data.
* EDA helps formulate hypotheses and decide on the most appropriate sampling methods.

**Feature Engineering:**

* Create new features or modify existing features to improve the predictive power of the data.
* Feature selection helps to reduce dimensionality and improve model performance.

**Modeling:**

* Depending on the nature of the problem (e.g., classification, regression, clustering), select and train machine learning or statistical models.
* Divide the data into training and testing sets for model analysis.
* Tune hyper parameters to improve model performance through techniques such as cross-validation.

**Model Evaluation:**

* Use various evaluation metrics (e.g., accuracy, precision, recall, F1-score, RMSE) to assess the model's performance.
* Use techniques such as cross-validation and holdout testing to ensure robustness.

**Model Deployment:**

* Once you have a satisfactory model, send it to a manufacturing facility for real-time forecasting or decision support.
* Consider things like scalability, monitoring, and integration with other systems.

**Monitoring and Maintenance:**

* Continuously monitor the performance of the model implemented to ensure that it remains accurate and relevant.
* Retrain the model periodically

**Visualization and Reporting**

* Create graphics and reports to communicate observations and findings to stakeholders.
* Visualization can help with storytelling and data-driven decision-making.

**Documentation:**

* Review appropriate documentation for all steps in the pipeline, including rules, data transformations, and model configurations.
* Documentation is essential for reproducibility and collaboration.

**Feedback Loop:**

* Collect feedback from end users and stakeholders to improve both the model and the pipeline.
* Iteratively refine and grow the pipeline based on feedback and changing business needs