Assignment-1

Q-1: Application of data science with example.

Ans:-

1. Recommendation Systems:

- **Application:** Personalizing user experiences by suggesting products, content, or services.
- **Example:** When you browse **Amazon**, it suggests "Customers who bought this also bought..." or "Recommended for you based on your Browse history." Similarly, **Netflix** recommends movies and TV shows based on your viewing habits, ratings, and those of similar users. This is achieved by analyzing vast amounts of user data, item attributes, and historical interactions using algorithms like collaborative filtering and content-based filtering.

2. Fraud Detection:

- **Application:** Identifying and preventing fraudulent activities in financial transactions, insurance claims, and other areas.
- **Example: PayPal** and credit card companies use data science to monitor millions of transactions in real-time. If a customer who usually makes small local purchases suddenly has a large overseas transaction, data science algorithms can flag it as potentially fraudulent, triggering an alert for investigation. This involves anomaly detection and pattern recognition techniques to identify unusual behavior.

3. Healthcare and Medicine:

- **Application:** Improving diagnosis, personalizing treatment, accelerating drug discovery, and optimizing hospital operations.
- **Example:**
    - **Identifying Cancer Tumors:** Google's **LYNA** (Lymph Node Assistant) tool uses machine learning to identify breast cancer tumors that metastasize to nearby lymph nodes with high accuracy (99% in trials).
    - **Personalized Treatment Plans:** Companies like **Oncora Medical** use machine learning to create personalized recommendations for cancer patients based on data from past cases, helping doctors make more efficient and tailored treatment decisions.
    - **Drug Discovery:** Data science can analyze vast datasets of molecular structures and biological information to predict the potential effectiveness and side effects of new compounds, significantly speeding up the drug development process.

4. Transportation and Logistics:

- **Application:** Optimizing routes, predicting traffic, improving delivery efficiency, and enhancing safety.
- **Example:**
    - **UberEats** uses data science to optimize food delivery routes, predicting factors like cooking time and traffic to ensure hot food is delivered quickly.
    - **UPS** uses its ORION (On-Road Integrated Optimization and Navigation) system, powered by data science, to optimize delivery routes for its drivers, saving millions of miles and gallons of fuel annually.
    - **Self-driving cars** heavily rely on data science to process sensor data, make real-time decisions, and navigate safely by analyzing vast amounts of training data (speed limits, road types, etc.).

5. Marketing and Advertising:

- **Application:** Targeted advertising, customer segmentation, market research, and sentiment analysis.
- **Example:** Companies use data science to analyze user search histories, Browse behaviors, and online activities to deliver **highly personalized advertisements** and product recommendations. If you search for a specific mobile phone online, you'll likely see ads for that phone across various websites and social media platforms. Sentiment analysis helps businesses gauge public opinion about their brand or products by analyzing social media and customer feedback.

6. Financial Services:

- **Application:** Risk assessment, algorithmic trading, credit scoring, and predicting market trends.
- **Example:** Banks use data science to assess the **creditworthiness** of individuals and businesses by analyzing their financial history, spending habits, and repayment behavior. data science models market data in real-time to execute trades automatically, aiming to maximize returns.

7. Education:

- **Application:** Adaptive learning, student performance analysis, curriculum improvement, and predicting student dropout rates.
- **Example:** Educational institutions can use data science to understand student performance, identify areas where students need more support, and even predict.

8. Manufacturing:

- **Application:** Predictive maintenance, supply chain optimization, and quality control.
- **Example:** In manufacturing, data science helps predict equipment failures before they occur (**predictive maintenance**) by analyzing sensor data and usage patterns, thus preventing costly downtime and reducing maintenance expenses.

Q-2: Differentiate between Data Science, AI, and ML.

Ans:-

Aspect	Data Science	Artificial Intelligence (AI)	Machine Learning (ML)
Scope	Broad: Combines statistics, programming, domain expertise, and visualization.	Broad: Encompasses ML, rule-based systems, robotics, NLP, etc.	Narrow: A subset of AI focused on data-driven learning.
Goal	Extract insights and solve problems using data.	Mimic human intelligence for autonomous task performance.	Learn from data to make predictions or decisions.
Techniques	Statistical analysis, data wrangling, visualization, ML, and domain knowledge.	ML, deep learning, NLP, computer vision, expert systems, robotics.	Supervised, unsupervised, reinforcement learning algorithms.
Tools	Python, R, SQL, Pandas, Tableau, Spark, ML libraries (e.g., scikit-learn).	TensorFlow, PyTorch, ROS, OpenCV, knowledge bases, Grok 3.	Scikit-learn, TensorFlow, PyTorch, XGBoost, LightGBM.
Data Dependency	Heavily data-driven; includes data prep and exploration.	May use data or predefined rules; not always data-dependent.	Relies entirely on data for training models.
Exaples	Customer segmentation,	Chatbots, self-driving cars,	Recommendation systems, fraud

	sales forecasting, data dashboards.	voice assistants (e.g., Grok 3).	detection, image classification.
Output	Insights, reports, models, or visualizations.	Autonomous systems or intelligent behavior.	Predictive models or classifications.

Q-3: Explain any five tools of DS widely used for resource.

Ans:-

1. **Python**

- **Purpose**: A versatile programming language used for data analysis, machine learning, and visualization.
- **Why Used**: Python's extensive libraries like Pandas (data manipulation), NumPy (numerical computing), Scikit-learn (machine learning), and Matplotlib/Seaborn (visualization) make it a go-to tool for data scientists. Its simplicity and community support enhance its accessibility for handling large datasets and building models.
- **Use Case**: Data cleaning, exploratory data analysis, and deploying machine learning models.

2. **R**

- **Purpose**: A programming language and environment designed for statistical computing and graphics.
- **Why Used**: R excels in statistical analysis and visualization, with packages like ggplot2 for advanced plotting and dplyr for data manipulation. It's widely used in academia and industries like finance and pharmaceuticals for statistical modeling.
- **Use Case**: Statistical analysis, data visualization, and predictive modeling.

3. **SQL (Structured Query Language)**

- **Purpose**: A language for managing and querying relational databases.
- **Why Used**: SQL is essential for extracting and manipulating data stored in databases like MySQL, PostgreSQL, or SQLite. It allows data scientists to datasets for making it critical for working with large-scale data.
- **Use Case**: Querying databases to extract customer data or sales records for analysis.

4. **Jupyter Notebook**

- **Purpose**: An open-source web application for creating and sharing documents with live code, visualizations, and narrative text.
- **Why Used**: Jupyter supports interactive coding in Python, R, and other languages, making it ideal for data exploration, prototyping, and sharing results. Its ability to combine code, visualizations, and explanations in one place is highly valued.
- **Use Case**: Exploratory data analysis, model prototyping, and creating reproducible research reports.

5. **Tableau**

- **Purpose**: A powerful data visualization and business intelligence tool.
- **Why Used**: Tableau enables data scientists to create interactive dashboards and visualizations without extensive coding. It connects to various data sources and is user-friendly for non-technical stakeholders to explore data insights.
- **Use Case**: Building dashboards for business metrics or visualizing trends in large datasets.

Q-4: List various websites available for downloading datasets.

Ans:-

1. Kaggle (kaggle.com)

- Offers a vast collection of datasets for data science and machine learning, with user-contributed and competition-related datasets.

- Covers topics like finance, health, social media, and more.

2. UCI Machine Learning Repository (archive.ics.uci.edu)

- A classic resource with hundreds of datasets for machine learning research.

- Includes datasets in domains like biology, physics, and social sciences.

3. Google Dataset Search (datasetsearch.research.google.com)

- A search engine for datasets across the web, indexing open datasets from various sources.

- Useful for finding niche or specialized data.

4. Data.gov (data.gov)

- The U.S. government's open data portal with datasets on topics like agriculture, health, education, and climate.

- Free and publicly accessible.

5. AWS Open Data Registry (registry.opendata.aws)

- Hosts datasets available through Amazon Web Services, including large-scale data like satellite imagery and genomics.

- Often used for big data projects.

6. World Bank Open Data (data.worldbank.org)

- Provides global development data, including economic, environmental, and social indicators.

- Ideal for economic and policy research.

7. European Data Portal (data.europa.eu)

- Offers open data from European Union institutions and member states.

- Covers areas like transport, environment, and public sector information.

8. Figshare(figshare.com)

- A repository for research data, including datasets from academic papers.

- Useful for accessing raw research data across disciplines.

9. Dryad (datadryad.org)

- Focuses on research data associated with scientific publications.

- Strong in biological and environmental sciences.

10. Zenodo (zenodo.org)

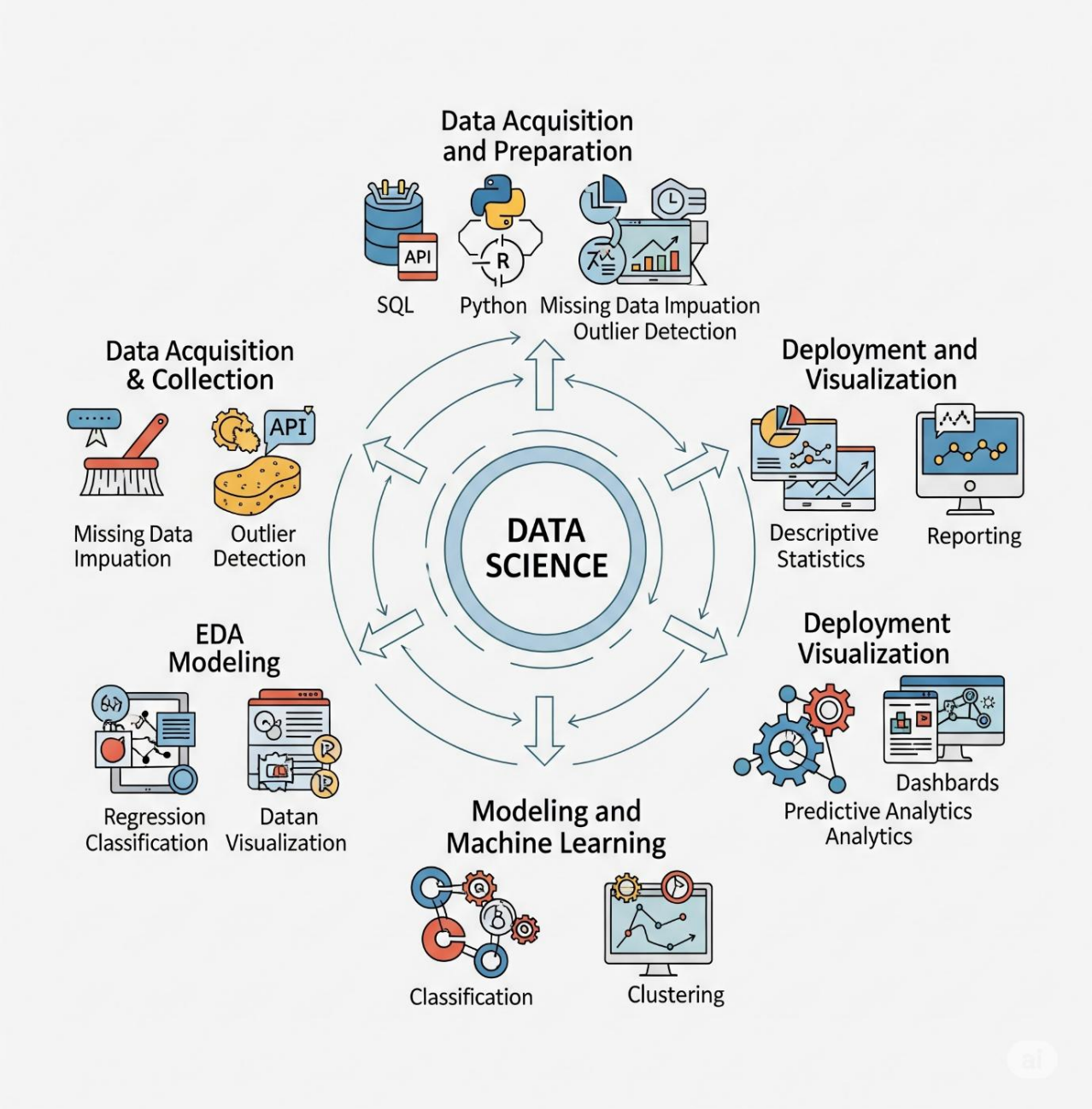
- A general-purpose open data repository for research data, funded by the European Commission.

- Supports datasets from any field with DOI assignment.

Q-5: Explain Data Science with a proper diagram.

Ans:-

Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It combines elements of statistics, computer science, and domain expertise to solve complex problems and make data-driven decisions.



The typical data science workflow involves several stages:

1. **Data Acquisition and Collection:** This initial phase involves gathering raw data from various sources. This can include databases, APIs, web scraping, sensors, and more. Tools and techniques like SQL, Python, and R are often used here.
2. **Data Cleaning and Preparation:** Raw data is often messy and inconsistent. This stage focuses on cleaning the data by handling missing values, correcting errors, removing duplicates, and transforming it into a suitable format for analysis.
3. **Exploratory Data Analysis (EDA):** Once the data is clean, EDA involves summarizing its main characteristics, often with visual methods. This helps in understanding patterns, detecting anomalies, and formulating hypotheses. Techniques include descriptive statistics and various data visualization methods.
4. **Modeling and Machine Learning:** In this stage, various statistical and machine learning models are applied to the prepared data. The goal is to build predictive models, classify data, or discover hidden patterns. This can involve techniques like regression, classification, clustering, and deep learning.
5. **Deployment and Visualization:** The final stage involves deploying the models into production systems and communicating the insights gained. This often includes creating interactive dashboards, reports, and web applications to help stakeholders understand and utilize the findings for decision-making.