BWT DATA SCIENCE TASK 1

SUB TOPIC 1: DEFINING DATA SCIENCE

What is Data?

- Data is everywhere in the form of text, phone numbers, time on a watch.
- We use data daily by counting money, writing letters.
- Computers need data to work.

What is Data Science?

- Data science extracts knowledge from data.
- Uses methods like probability and statistics.
- Applies knowledge to real-world problems.
- Handles both structured and unstructured data.
- Needs expertise in fields like finance or medicine.

Related Fields

- Databases: Store and structure data.
- Big Data: Process large data sets efficiently.
- Machine Learning: Build models to predict outcomes.
- AI: Mimic human thought with data.
- Visualization: Make data understandable through visuals.

Types of Data

- Structured: Organized in tables (e.g., phone lists).
- Unstructured: Raw files (e.g., videos).
- Semi-structured: Mixed form (e.g., web pages).

Where to Get Data

- Structured: IoT sensors, surveys.
- Unstructured: Texts, videos.
- Semi-structured: Social network graphs.

What You Can Do with Data

- Collect data.
- Store data efficiently.
- Convert data for use.
- Create visuals to understand data.
- Predict outcomes with machine learning.

Digitalization and Transformation

- Businesses use data for decisions.
- Convert processes to digital.
- Use data to improve productivity.
- Example: Improve online courses by analyzing student data.

SUB TOPIC 2: DATA SCIENCE ETHICS

Data Ethics and Trends

- AI makes app integration easier but has risks like misuse.
- By 2025, we will create and use over 180 zettabytes of data.
- Data scientists have access to personal data, influencing user behavior but raising privacy concerns.
- Data Ethics is necessary to prevent harm and misuse.
- Key Trends: Digital ethics, responsible AI, AI governance.

Ethics Basics

- Ethics are moral principles guiding behavior.
- Data Ethics evaluates moral issues related to data and algorithms.
- Applied Ethics relates to practical application of moral principles.
- Ethics Culture ensures ethical practices are consistently followed in organizations.

Ethics Principles

- Ethics principles means shared values to guide AI and data projects.
- Example: Microsoft's six principles include accountability, transparency, fairness, reliability, privacy, and inclusiveness.

Ethics Challenges

- Data Ownership: Who controls and owns the data?
- Consent: Did users understand and agree to data use?
- Data Privacy: Securing user data.
- Allowing data deletion on request.
- Dataset Bias: Ensure data represents all groups fairly.
- Data Quality: Validating data accuracy and consistency.
- Misrepresentation of data

Case Studies

Real-world examples show the impact of ethical issues:

- Informed Consent: Tuskegee Syphilis Study.
- Data Privacy: Netflix data re-identification.
- Collection Bias: Boston's Street Bump app.
- Algorithmic Fairness: MIT Gender Shades Study.
- Misrepresentation: Georgia COVID-19 data charts.

Applied Ethics Solutions

- Guidelines for ethical behavior in organizations.
- Practical tools for ensuring ethical practices.
- Laws like GDPR protect user data rights.
- Building a culture of ethics within organizations.

SUB TOPIC 3: DEFINING DATA

Data and Its Characteristics

- Data: Facts, information, observations, and measurements for discoveries and decisions.
- Data Point is a single unit of data in a dataset.
- Dataset is a collection of data points whose formats may vary (e.g., spreadsheet, JSON).

Types of Data

- Raw Data: It is unanalyzed from the source and needs organizing.
- Quantitative Data: Numerical, measurable (e.g., population, height).
- Qualitative Data: Categorical, not measured mathematically (e.g., comments, car models).

Data Structures

- Structured Data: Organized in rows and columns (e.g, spreadsheets, databases).
- Unstructured Data: No specific format (e.g, text files, videos).
- Semi-structured Data: Combination of both structured and unstructured data (e.g, HTML, JSON).

Sources of Data

- Primary Data: Collected directly by users (e.g, scientists' observations).
- Secondary Data: Collected for general use, shared with others.

Common Data Sources

- Databases: Managed systems for data storage.
- Files: Audio, image, video, spreadsheets.
- Internet: Hosts data, accessible via APIs and web scraping.