Data Science SW - Jaish Khan

Table of Contents

for BSSW-P3-Morning.

---

1. Data

> Data is the raw, unorganized facts, figures, and symbols. It is the basic building block of information and can exist in various forms, such as numbers, text, images, audio, and video

Variables → characteristics or properties within a dataset that can vary. In a student dataset, variables could include Age, Name, Gender, and Marks.

- *Independent Variables* (Features): Variables that stand alone and can influence other variables. Example: Number of study hours.
- *Dependent Variables* (Target): Variables that depend on other variables. Example: Final exam score (depends on study hours).

1.1. Data Types

> They are fundamental categories that define how data should be handled and processed.

Why Data Types Matter

Understanding data types is important because they:

- Determine how data is stored and processed
- Affect memory usage and computational efficiency
- Influence the choice of analysis methods
- Impact the accuracy of results

KEY DATA TYPES

1. **Numerical Data** (Numbers)
   - *Integers* → Whole numbers (Counting data, Age in years etc.)
   - *Floats* → Decimal numbers (Measurements, Percentages, Financial data etc.)
2. **Categorical Data** (Categories)
   - *Nominal* → Categories without order (Colors, Gender, Blood Types etc.)
   - *Ordinal* → Categories with order (Rank, Satisfaction, Grades etc.)
3. **Text Data** (Text)
   - Used for → Social media posts, Product descriptions, Email content etc.
4. **Boolean Data** (Binary)
   - Used for → True/False flags, Status indicators, Condition checks etc.

1.2. Structured vs. Unstructured Data

Structured Data → Organized in predefined formats, Follows a schema or data model and it is easy to search and analyze.

- *Examples*: SQL databases, Excel spreadsheets, CSV files, Time series data

Unstructured Data → No predefined structure, Harder to process and analyze and requires specialized tools.

- *Examples*: Text documents, Images and videos, Audio files, Social media content

> **Working with Different Data Types**
>
> When working with different data types:
>
> 1. Always validate data types before analysis
> 2. Convert data types when necessary
> 3. Handle missing values appropriately
> 4. Document any data type transformations

1.3. Data Collection Methods

Primary data collection involves gathering data directly from the source.

- *Advantages*: The data is tailored to your specific needs. Primary data is generally considered to be more accurate.
- *Disadvantages*: Collecting primary data can be time-consuming. Primary data collection can be expensive.
- *Examples*: surveys, interviews, and observations.

Secondary data collection uses data that has already been collected by someone else.

- *Advantages*: Secondary data is usually cost-effective. Secondary data is readily available.
- *Disadvantages*: Secondary data may not perfectly fit your needs. The data may be outdated.
- *Examples*: government databases, research papers, and online sources.

> **Choosing Collection Methods**
>
> The choice of data collection method depends on:

- Research objectives
- Available resources
- Time constraints
- Data quality requirements

1.4. Data Quality Characteristics

1. *Accuracy* → Data correctness, Minimal errors, Verified sources
2. *Completeness* → No missing values, All required fields present, Comprehensive coverage
3. *Consistency* → Uniform formats, Standardized values, No contradictions
4. *Timeline* → Up-to-date information, Regular updates, Relevant time period

1.5. Common Data Issues

⚠ **Data Quality Issues**

Common problems that need addressing:

1. **Missing Data**

   ☰ **Handling Missing Data**
   - **Detection Methods** → Null value checks, Empty string checks, Special value checks (e.g., -999)
   - **Solutions** → Remove rows (if few missing values), Impute with mean/median, Use advanced imputation methods, Create 'missing' category

2. **Duplicate Data**

   ☰ **Handling Duplicate Data**
   - **Identification** → Exact matches, Fuzzy matches, Business key matches
   - **Solutions** → Remove duplicates, Merge information, Keep most recent, Flag duplicates

3. **Inconsistent Data**

   ☰ **Handling Inconsistent Data**
   - **Common Issues** → Different date formats, Varying units, Inconsistent spelling

- **Solutions** → Format standardization, Unit conversion, Spelling normalization

1.6. Data Transformation Techniques

1. **Normalization** → A data preprocessing to adjust numerical values to a common scale, typically between 0 and 1. This process prevents variables with large ranges from dominating the analysis and ensures that all features contribute equally to the machine learning model.
   - *Min-Max Scaling* → Formula: $(x - min)/(max - min)$ | Range: $[0, 1]$
   - *Z-Score Normalization* → Formula: $(x - mean)/std$ | Range: $(-3, 3)$
   - *Decimal Scaling* → Formula: $x/10^n$ | Range: varies
2. **Encoding** → A data preprocessing technique that converts text-based categorical variables into a numerical format that machine learning algorithms can understand.
   - *One-Hot Encoding* → Binary columns for each category, No ordinal relationship
   - *Label Encoding* → Single column with integers, Maintains ordinal relationship
   - *Binary Encoding* → Binary representation, Memory efficient

---

2. Exploratory Data Analysis (EDA)

Purpose of EDA

EDA helps:

- Understand data distribution
- Identify patterns and relationships
- Detect anomalies
- Generate hypotheses

2.1. Summary Statistics

1. Measures of **Central Tendency**
   - *Mean*: arithmetic average
   - *Median*: middle value
   - *Mode*: most frequent value
2. Measures of **Spread**
   - *Range*: max value - min value
   - *Variance*: average squared deviation
   - *Standard deviation*: square root of variance
   - *Quartiles*: divide data into four parts
3. Measures of **Shape**
   - *Skewness*: asymmetry measure
   - *Kurtosis*: tail heaviness measure

2.2. Visualization Techniques

1. Univariate or Single Variable Plots
   1. **Histograms** → Show distribution, Identify patterns, Detect outliers
   2. **Box Plots** → Show quartiles, Identify outliers, Compare groups
   3. **Bar Charts** → Compare categories, Show frequencies, Display proportions
2. Bivariate or Two Variable Plots
   1. **Scatter Plots** → Show relationships, Identify correlations, Detect patterns
   2. **Line Plots** → Show trends, Compare changes, Time series analysis
   3. **Heat Maps** → Show correlations, Identify patterns, Compare categories

3. Machine Learning

> Machine learning is a subset of artificial intelligence (AI) that allows systems to learn from data, improving their performance over time without explicit programming.

Machine learning *enables computers* to: Learn from data, Identify patterns, Make predictions and Improve with experience.

Uses of Machine Learning

- **Predictive analytics:** Forecasting future trends, such as sales forecasting.
- **Recommendation systems:** Suggesting products or services based on user preferences, as seen on Netflix and Amazon.
- **Image and speech recognition:** Identifying faces or understanding spoken language, used in facial recognition software and voice assistants.
- **Natural language processing:** Enabling computers to understand and process human language, used in chatbots and language translation tools.

3.1. Types of Machine Learning

1. **Supervised Learning:** The model is trained on labeled data (which includes input-output pairs) This means the algorithm is given both the input features and the desired output, allowing it to learn the relationship between them.
   - *Types* → **Classification** and **Regression**
   - *Algorithms* → Linear Regression, Decision Trees, Support Vector Machines (SVM) etc.
   - *Uses* → Email Spam Detection, Credit Scoring etc.
2. **Unsupervised Learning:** The model is trained on unlabeled data, meaning it must discover patterns and groupings on its own without specific output guidance.
   - *Types* → **Clustering**, **Association** and **Dimensionality Reduction**
   - *Algorithms* → K-Means Clustering, Principal Component Analysis (PCA), and Hierarchical Clustering.
   - *Uses* → Pattern discovery, Feature extraction etc.
3. **Reinforcement Learning:** This type involves an agent learning to make decisions in an environment to maximize cumulative rewards. The agent interacts with the environment, receives feedback in the form of rewards or penalties, and adjusts its actions accordingly.
   - *Algorithms* → Q-Learning and Deep Q-Networks (DQN).
   - *Uses* → Games, Robots, Trading strategies etc.

3.2. Machine Learning Workflow

1. *Problem Definition* → Clearly define the problem you want to solve.
2. *Data Collection* → Gather relevant data from various sources.
3. *Data Preprocessing* → Prepare the data for analysis by cleaning, transforming, and handling missing values.
4. *Model Selection* → Choose an appropriate machine learning algorithm based on the problem type and the nature of the data.
5. *Training the Model* → Use the prepared data to train the model, allowing it to learn patterns and relationships within the data.
6. *Model Evaluation* → Evaluate the model's performance using metrics such as accuracy, precision, recall, and F1-score.
7. *Model Deployment* → Once the model is trained and evaluated, it can be deployed for real-world use, making predictions or decisions based on new data.

3.3. Neural Networks

> A specific type of machine learning model inspired by the structure and function of the human brain. They consist of interconnected nodes, or "neurons," organized in layers.

Neuron Structure → Inputs, Weights, Bias, Activation function, Output.

Layered Structure:

1. *Input Layer*: Receives the raw data input, similar to how our senses gather information.
2. *Hidden Layers*: Process the information from the input layer, extracting features and patterns through multiple layers of interconnected neurons.
3. *Output Layer*: Provides the final prediction or decision based on the processed information.

Learning Process:

1. *Weights and Biases*: Each connection between neurons has an associated weight, representing its importance. Biases are additional parameters that help fine-tune the network's learning.
2. *Forward Propagation*: The input data flows through the network, with each layer processing and transforming the information until it reaches the output layer, producing a prediction.
3. *Backpropagation*: The network compares its prediction to the actual output and adjusts the weights and biases based on the error, improving its accuracy over time.