جامـــــعة المـــــستقبل
AL MUSTAQBAL UNIVERSITY

قســـم الامـــــن الـــــــسيبرانـــــــي

Department of Cyber Security

Subject:

Data Wrangling

Class:

first

Lecturer:

Asst. lect. Mustafa Ameer Awadh

Lecture: (13)

Introduction

Data wrangling is the process of cleaning, structuring, and enriching raw data into a desired format for better analysis. It is a crucial step in the data science workflow to ensure data quality and consistency.

---

2. Steps in Data Wrangling

a. Data Collection

- Gathering data from various sources such as databases, APIs, and web scraping.
- Understanding the nature of structured vs. unstructured data.

b. Data Cleaning

- Handling missing values: Imputation techniques (mean, median, mode) and removing missing data.
- Removing duplicate records and inconsistencies.
- Handling outliers using statistical methods.

c. Data Transformation

- Normalization and standardization.
- Encoding categorical variables (One-Hot Encoding, Label Encoding).
- Feature scaling (Min-Max Scaling, Z-score normalization).

d. Data Integration

- Merging multiple datasets.
- Resolving data conflicts and inconsistencies.

e. Data Reduction

- Feature selection techniques to retain the most relevant attributes.
- Principal Component Analysis (PCA) for dimensionality reduction.

3. Tools for Data Wrangling

- **Python Libraries**: Pandas, NumPy, OpenRefine.
- **SQL for data querying**.
- **Cloud-based services**: AWS Glue, Google DataPrep.

4. Importance of Data Preprocessing

- Enhances model performance by ensuring high-quality data.
- Reduces computational costs by removing redundant data.
- Improves interpretability of the dataset for analysis.

5. Practical Example

Scenario: You have a dataset containing customer transactions. Steps:

1. Identify and handle missing values.
2. Convert categorical columns into numerical format.
3. Scale numerical features appropriately.
4. Merge with additional customer demographic data.

Conclusion

Data wrangling and preprocessing are fundamental to any data-driven project. Proper handling of data ensures accurate and meaningful insights, making analysis more efficient and reliable.