Logistic Regression - Student Learning Guide

1. Introduction

Data cleaning and preprocessing are essential first steps before training any machine learning model.

Poor quality data leads to poor quality predictions. This guide walks through best practices to prepare data effectively.

2. Handling Missing Values

Handling Missing Values:

- Identify missing data using df.isnull().sum()
- Drop rows/columns (if sparse or unimportant)
- Impute with mean, median, mode, or predictive models
- Consider domain-specific logic (e.g., fill 'age' by group median)

3. Fixing Data Types

Fixing Data Types:

- Ensure each column has the correct data type: int, float, object, datetime
- Convert columns using df.astype() or pd.to_datetime()
- Helps avoid silent errors and speeds up computation

4. Encoding Categorical Variables

Encoding Categorical Variables:

- Label Encoding: Turns categories into numeric codes (Ordinal)
- One-Hot Encoding: Creates binary columns for each category (Nominal)
- Avoid high-cardinality (too many unique values) when possible

5. Detecting and Handling Outliers

Detecting and Handling Outliers:

- Use boxplots, Z-scores, or IQR to detect outliers
- Remove, clip, or transform (e.g., log) depending on context

- Don't remove blindly - investigate why they exist

6. Removing Duplicates

Removing Duplicates:

- Check with df.duplicated() and remove using df.drop_duplicates()
- Especially useful in data collection or merging processes

7. Normalizing and Scaling

Normalizing and Scaling Features:

- StandardScaler: zero mean, unit variance (Z-score)
- MinMaxScaler: scales data to range [0, 1]
- Required for most ML models (except tree-based models)

8. Final Pre-Modeling Checks

Final Checks Before Modeling:

- Ensure no NaNs or infinite values remain
- All features must be numeric (convert categorical!)
- Target variable must be correctly encoded (binary or multiclass)