PROJECT TITLE

Healthcare Data Analysis for Disease Risk Prediction and Patient Outcome Improvement

Internship Role: Data Analyst
Name: Keerti Neeramanigar
Week: Week 2 – Data Cleaning and Preprocessing Methodology

Project Title

Healthcare Data Analysis for Disease Risk Prediction and Patient Outcome Improvement

1. Introduction

In any healthcare data analysis project, data cleaning and preprocessing is a very important step. Healthcare data often contains missing values, wrong entries, duplicate records, and inconsistent formats. If data is not cleaned properly, the analysis results may be incorrect and misleading.

This document explains a simple and systematic approach to clean and preprocess healthcare data before analysis. The methodology is explained using simple descriptions and Python-related terms. No actual code execution is done. The goal is to prepare clean, accurate, and reliable data for further analysis and visualization.

2. Initial Data Quality Assessment

Before cleaning the data, the first step is to understand the dataset.

The following checks are performed:

- Check number of rows and columns
- Understand column names and data types
- Identify missing values
- Identify duplicate records
- Check for unusual or extreme values (outliers)

In Python, libraries like **Pandas** are useful to inspect the dataset structure and basic information.

3. Common Data Quality Issues in Healthcare Data

Healthcare datasets usually face these problems:

- Missing values in patient age, test results, or diagnosis
- Duplicate patient records
- Outliers such as very high or very low values
- Inconsistent categorical values (e.g., “Male”, “male”, “M”)
- Different scales in numerical data

Identifying these issues early helps in choosing the correct cleaning method.

4. Data Cleaning and Preprocessing Methodology

4.1 Handling Missing Values

Missing values are common in healthcare data.

Methods used:

- If missing values are small → replace with mean or median
- If missing values are large → remove the column or row
- For categorical data → replace with most frequent value

This helps in maintaining data completeness.

4.2 Removing Duplicate Records

Duplicate records can affect analysis results.

Steps:

- Identify duplicate rows
- Remove repeated entries
- Keep only unique patient records

This ensures each patient is counted only once.

4.3 Handling Outliers

Outliers are values that are very different from others.

Examples:

- Very high blood pressure value
- Unusual age values

Methods:

- Detect outliers using statistical methods
- Remove or cap extreme values

This prevents incorrect influence on analysis.

4.4 Handling Categorical Data

Categorical data includes gender, disease type, or treatment type.

Steps:

- Convert text to consistent format (lowercase)
- Fix spelling differences
- Encode categories into numbers if needed

This helps Python tools process the data correctly.

4.5 Data Normalization

Normalization is used to bring numerical values into a similar range.

Example:

- Age (0–100)
- Blood sugar (70–200)

Normalization helps in fair comparison and better model performance.

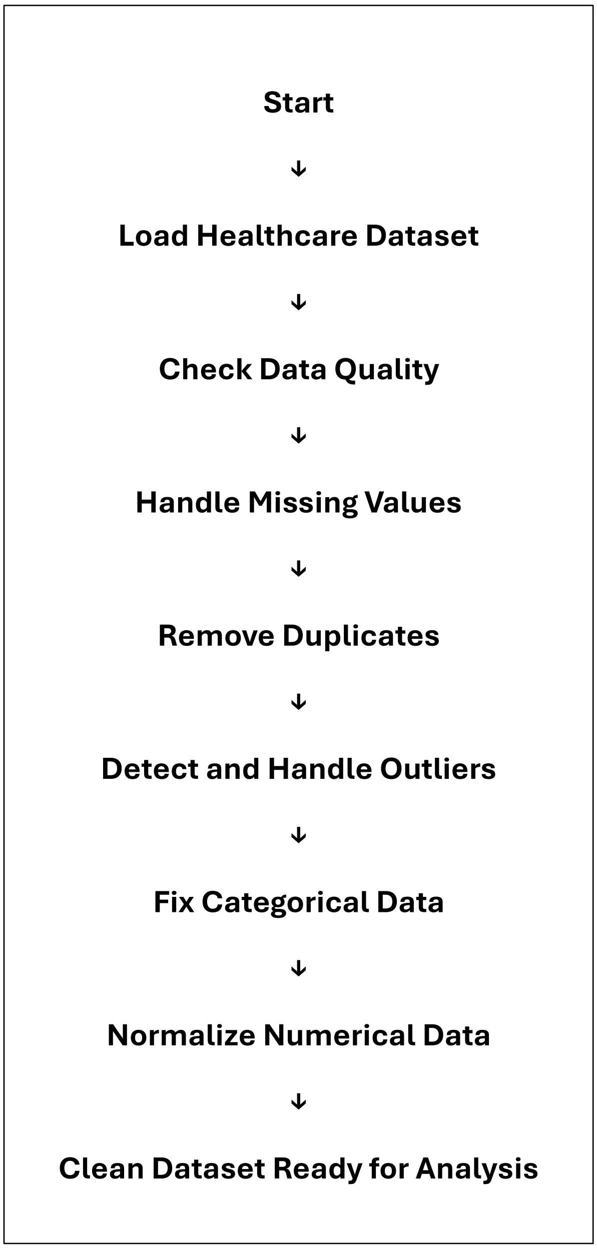
5. Tools and Libraries Used

The following Python libraries are useful in data cleaning:

- **Pandas** – data loading, cleaning, and manipulation
- **NumPy** – numerical operations and calculations

These libraries make data preprocessing simple and efficient.

6. Data Cleaning Workflow (Flowchart)



7. Summary Table of Data Cleaning Steps

Step	**Action Taken**	**Purpose**
Missing values	Imputation / removal	Improve data completeness
Duplicates	Remove repeated records	Avoid double counting
Outliers	Detect and adjust	Prevent wrong analysis
Categorical data	Standardize values	Maintain consistency
Normalization	Scale numerical values	Fair comparison

8. Conclusion

This document explains a simple and clear approach to data cleaning and preprocessing for healthcare datasets. Proper cleaning ensures accurate analysis and meaningful insights. By handling missing values, duplicates, outliers, and data inconsistencies, the dataset becomes reliable and ready for further analysis. This methodology can be applied to real-world healthcare data projects to ensure data quality and integrity.