Title: Healthcare Data Cleaning: Improving Disease Prediction Accuracy by Handling Missing, Inconsistent, and Noisy Patient Data
NAME = ADITYA TYAGI

ROLL NO = 202401100400018
Date: [10/3/2025]
Institution: [KIET group of institutions]

---

2. Introduction

In the healthcare industry, accurate disease prediction is crucial for effective patient care and decision-making. However, healthcare data often comes with issues such as missing values, inconsistencies, and noise, which can significantly affect the accuracy of disease prediction models. This project aims to address these challenges by using various data cleaning techniques to improve the quality of patient data. Through the process of cleaning the data, we will be able to train more reliable machine learning models that can predict diseases more accurately, leading to better health outcomes.

The objective of this project is to demonstrate the importance of data cleaning and provide solutions to handle missing, inconsistent, and noisy data. This will be achieved by exploring various methods of data imputation, outlier detection, and noise filtering.

---

3. Methodology

The project follows a structured approach for cleaning and preparing the data, which can be broken down into the following key steps:

3.1. Data Collection

The dataset used in this project consists of patient information such as age, gender, medical history, test results, and diagnosis. These datasets are typically collected from hospitals or healthcare centers and contain various issues that need to be addressed.

3.2. Identifying Missing Data

Handling missing data is crucial because it can affect the performance of machine learning algorithms. We used the following techniques to deal with missing values:

- **Deletion Method**: Rows or columns with a high percentage of missing values were removed.
- **Imputation Method**: For numerical columns, we used the mean or median to impute missing values. For categorical columns, the mode was used for imputation.

3.3. Handling Inconsistent Data

Inconsistent data refers to entries that don't conform to expected formats or values. For example, a numeric field like "age" might have text values or unrealistic numbers. To handle this:

- **Standardization**: We ensured that numeric fields were in the correct range, and categorical values followed a consistent naming convention.
- **Conversion**: Any incorrectly formatted entries were converted to the correct format (e.g., converting "Male" and "M" to "M").

3.4. Noise Filtering

Noise in data refers to random errors or variations that do not contribute to the predictive power of the model. To reduce noise:

- **Outlier Detection**: We identified outliers using statistical methods such as the Z-score and IQR (Interquartile Range) methods. These outliers were either corrected or removed.
- **Smoothing**: Techniques like moving averages were applied to reduce random fluctuations in data.

3.5. Data Normalization

Once the data was cleaned, we applied normalization to scale numerical values to a similar range. This is important because machine learning models can perform better when features are on a similar scale.