Report: Healthcare Data Cleaning
Title Page
Problem Statement:
Healthcare Data Cleaning: Improve disease prediction accuracy by handling missing, inconsistent, and noisy patient data.

Personal Details:

Name: Nischhal Garg

Roll No.: 202401100400131

Introduction
In healthcare, accurate disease prediction is crucial for effective treatment and patient care. However, real-world healthcare datasets often contain missing, inconsistent, and noisy data, which can significantly impact the performance of predictive models. This project focuses on cleaning a synthetic healthcare dataset to improve the accuracy of disease prediction. The dataset includes patient information such as age, blood pressure, cholesterol levels, and disease labels.

The goal of this project is to:

Handle missing data by imputing or removing it.

Fix inconsistencies in categorical data (e.g., standardizing disease names).

Remove noisy data (e.g., outliers in numerical columns).

Prepare the cleaned dataset for further analysis or machine learning modeling.

Methodology
The following steps were taken to clean the dataset and improve disease prediction accuracy:

Generate Synthetic Dataset:

A synthetic healthcare dataset was generated with 200 records, including patient ID, age, blood pressure, cholesterol, and disease labels.

Missing values were introduced in the blood pressure and cholesterol columns.

Inconsistencies were introduced in the disease column (e.g., misspelled disease names).

Handle Missing Data:

Missing values in the blood pressure column were filled with the mean value.

Missing values in the cholesterol column were filled with the median value.

Fix Inconsistencies:

Disease names were standardized by converting them to lowercase and correcting misspelled entries (e.g., "Hpertension" to "Hypertension").

Disease names were capitalized for consistency.

Remove Outliers:

Outliers in the numerical columns (age, blood pressure, cholesterol) were removed using the Z-score method.

Visualize Data:

A boxplot was created to visualize the distribution of the numerical columns and identify outliers.

Save Cleaned Dataset:

The cleaned dataset was saved to a CSV file for further use.

Code

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Generate synthetic healthcare dataset
def generate_healthcare_data(num_records=200):
    np.random.seed(42)

    # Generate synthetic data
    patient_ids = np.arange(1, num_records + 1)
    ages = np.random.randint(20, 90, num_records)
    blood_pressure = np.random.normal(120, 15, num_records)
    blood_pressure[np.random.choice(num_records, 10, replace=False)] = np.nan  # Inject missing values
    cholesterol = np.random.normal(200, 40, num_records)
    cholesterol[np.random.choice(num_records, 5, replace=False)] = np.nan  # Inject missing values
    disease = np.random.choice(['Diabetes', 'Hypertension', 'Heart Disease', 'None'], num_records)
    disease[5] = 'Hpertension'  # Introduce inconsistency
    disease[10] = 'diabetes'     # Introduce inconsistency

    data = pd.DataFrame({
        'Patient_ID': patient_ids,
        'Age': ages,
        'Blood_Pressure': blood_pressure,
        'Cholesterol': cholesterol,
        'Disease': disease
    })

    return data

# Handle missing values
def handle_missing_values(df):
    df = df.copy()
    df.loc[:, 'Blood_Pressure'] = df['Blood_Pressure'].fillna(df['Blood_Pressure'].mean())
    df.loc[:, 'Cholesterol'] = df['Cholesterol'].fillna(df['Cholesterol'].median())
    return df

# Fix inconsistencies
def standardize_disease_names(df):
    df = df.copy()
    df.loc[:, 'Disease'] = df['Disease'].str.lower().str.replace('hpertension', 'hypertension')
    df.loc[:, 'Disease'] = df['Disease'].str.capitalize()
    return df

# Remove outliers (Z-score method)
def remove_outliers(df):
    from scipy.stats import zscore
```

```
    df = df[np.abs(zscore(df[['Age', 'Blood_Pressure', 'Cholesterol']])) < 3].all(axis=1)
    return df

# Visualize data distribution
def visualize_data(df):
    plt.figure(figsize=(12, 5))
    sns.boxplot(data=df[['Age', 'Blood_Pressure', 'Cholesterol']])
    plt.title("Boxplot for Outlier Detection")
    plt.show()

# Main execution
healthcare_data = generate_healthcare_data()
healthcare_data = handle_missing_values(healthcare_data)
healthcare_data = standardize_disease_names(healthcare_data)
visualize_data(healthcare_data)
healthcare_data = remove_outliers(healthcare_data)

# Save cleaned dataset
healthcare_data.to_csv("cleaned_healthcare_data.csv", index=False)
print("Cleaned dataset saved as cleaned_healthcare_data.csv")
```

References/Credits

Dataset: Synthetic dataset generated using Python.

ouput

Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scipy.

Techniques: Imputation, Outlier Removal, Z-score method.

Conclusion

This project successfully cleaned a synthetic healthcare dataset by handling missing, inconsistent, and noisy data. The cleaned dataset is now ready for further analysis or machine learning modeling. The steps taken in this project can be applied to real-world healthcare datasets to improve disease prediction accuracy.