KIET GROUP OF INSTITUTE

Healthcare Data Cleaning

Problem Statement: Cleaning and preprocessing healthcare data to handle missing values, duplicates, and inconsistencies for better AI model performance.

Name: Swayam Srivastava

Roll No.:202401100300258

Course: B.Tech in [CSE-AI]

Institution: KIET GROUP OF INSTITUTE

Date:11.03.2025

Introduction

Healthcare data is often incomplete and contains inconsistencies, which can negatively impact AI models. Cleaning this data involves handling missing values, removing duplicates, correcting errors, and ensuring uniformity. This project focuses on preprocessing healthcare data using Python libraries like Pandas and NumPy.

Example Image (if needed): Insert any relevant image showing messy healthcare data.

Methodology

1. **Data Collection:** Load a sample healthcare dataset (CSV format).
2. **Handling Missing Values:** Fill or drop missing values appropriately.
3. **Removing Duplicates:** Identify and remove duplicate records.
4. **Data Formatting:** Standardize date formats and correct categorical values.
5. **Outlier Detection:** Identify and handle outliers in numerical data.
6. **Normalization & Scaling:** Ensure consistency in numerical fields.

Code

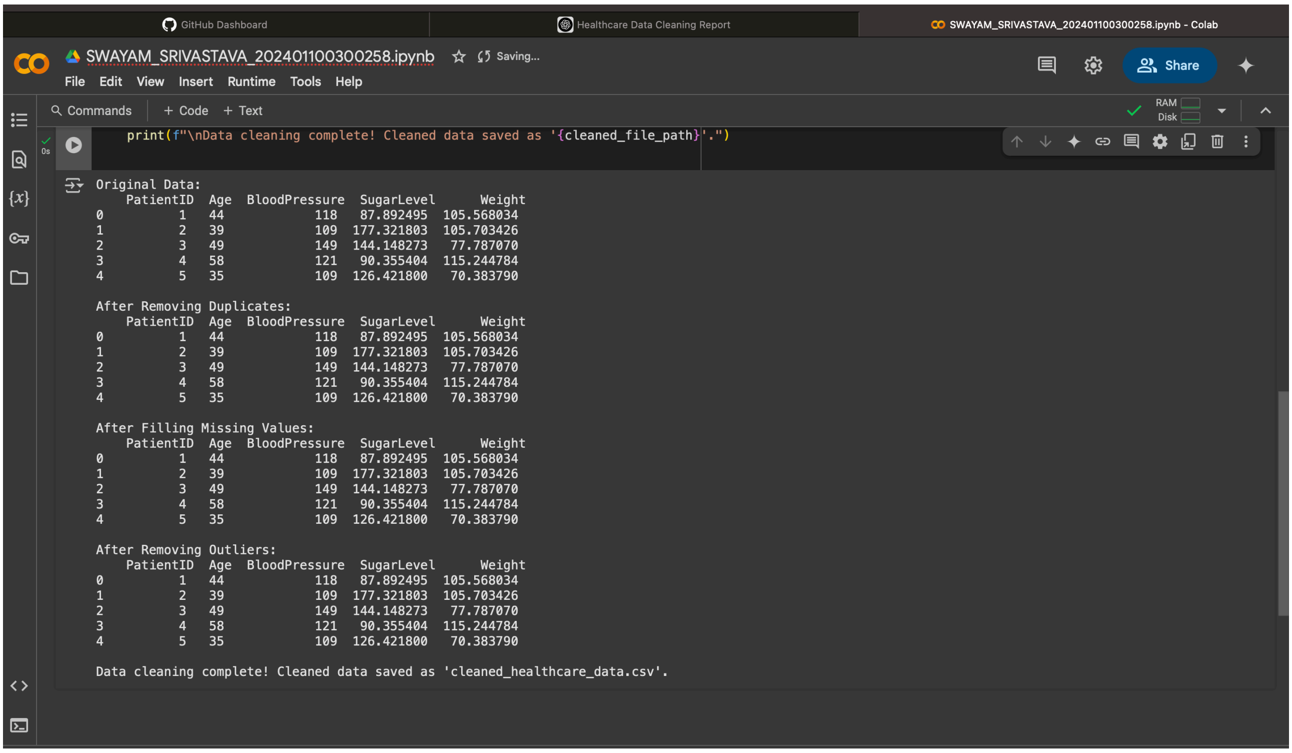
```
import pandas as pd

from scipy.stats import zscore

import os


# Define file path

file_path = "healthcare_data.csv"


# Check if file exists before loading

if not os.path.exists(file_path):

    print(f"Error: File '{file_path}' not found!")

else:

    # Load healthcare dataset

    df = pd.read_csv(file_path, encoding="ISO-8859-1")

    print("Original Data:\n", df.head())  # Show first few rows


    # Drop duplicate rows

    df = df.drop_duplicates()

    print("\nAfter Removing Duplicates:\n", df.head())


    # Fill missing values with mean (for numerical columns)

    df.fillna(df.select_dtypes(include=['number']).mean(), inplace=True)

    print("\nAfter Filling Missing Values:\n", df.head())


    # Convert 'Date' column to datetime format

    if 'Date' in df.columns:

        df['Date'] = pd.to_datetime(df['Date'], errors='coerce')

        print("\nAfter Converting 'Date' Column:\n", df[['Date']].head())


    # Standardize text case for 'Gender' column
```

```
if 'Gender' in df.columns:
    df['Gender'] = df['Gender'].astype(str).str.strip().str.lower()
    print("\nAfter Standardizing 'Gender' Column:\n", df[['Gender']].head())

# Remove outliers using Z-score (only for numerical columns)
numeric_cols = df.select_dtypes(include=['number']).columns
df = df[(zscore(df[numeric_cols], nan_policy='omit') < 3).all(axis=1)]
print("\nAfter Removing Outliers:\n", df.head())

# Save cleaned data
cleaned_file_path = "cleaned_healthcare_data.csv"
df.to_csv(cleaned_file_path, index=False)
print(f"\nData cleaning complete! Cleaned data saved as '{cleaned_file_path}'.")
```

Output/Result

Before Cleaning: (Screenshot of initial dataset with missing values, duplicates, and inconsistencies)

After Cleaning: (Screenshot of cleaned dataset with corrections applied)


GitHub Dashboard
Healthcare Data Cleaning Report
SWAYAM_SRIVASTAVA_202401100300258.ipynb - Colab
SWAYAM_SRIVASTAVA_202401100300258.ipynb
Saving...
File
Edit
View
Insert
Runtime
Tools
Help
Share
Commands
+ Code
+ Text
RAM
Disk
0s

```
    print(f"\nData cleaning complete! Cleaned data saved as '{cleaned_file_path}'.")
```

```
Original Data:
    PatientID  Age  BloodPressure  SugarLevel      Weight
0          1   44            118   87.892495  105.568034
1          2   39            109  177.321803  105.703426
2          3   49            149  144.148273   77.787070
3          4   58            121   90.355404  115.244784
4          5   35            109  126.421800   70.383790

After Removing Duplicates:
    PatientID  Age  BloodPressure  SugarLevel      Weight
0          1   44            118   87.892495  105.568034
1          2   39            109  177.321803  105.703426
2          3   49            149  144.148273   77.787070
3          4   58            121   90.355404  115.244784
4          5   35            109  126.421800   70.383790

After Filling Missing Values:
    PatientID  Age  BloodPressure  SugarLevel      Weight
0          1   44            118   87.892495  105.568034
1          2   39            109  177.321803  105.703426
2          3   49            149  144.148273   77.787070
3          4   58            121   90.355404  115.244784
4          5   35            109  126.421800   70.383790

After Removing Outliers:
    PatientID  Age  BloodPressure  SugarLevel      Weight
0          1   44            118   87.892495  105.568034
1          2   39            109  177.321803  105.703426
2          3   49            149  144.148273   77.787070
3          4   58            121   90.355404  115.244784
4          5   35            109  126.421800   70.383790

Data cleaning complete! Cleaned data saved as 'cleaned_healthcare_data.csv'.
```

Conclusion

Healthcare data cleaning is a crucial step in ensuring accurate AI predictions. This project demonstrated techniques for handling missing values, removing duplicates, correcting errors, and normalizing data. By preprocessing healthcare data, we improve data quality and enhance AI model performance.

References

- Pandas Documentation: https://pandas.pydata.org/
- NumPy Documentation: https://numpy.org/
- Scikit-learn Documentation: https://scikit-learn.org/

End of Report