Data Preprocessing Pipeline for Disease Incidence Analysis

Sainath R.

June 13, 2025

Overview

- Objective: Preprocess raw dataset to enable disease incidence analysis.
- Focused on normalization, missing value imputation, outlier treatment, and dimensionality reduction.
- Key columns include:
  - `Incidence of TB`, `Hypertension`, `Diabetes Prevalence`
  - `Start of Screening coverage (year)`, `Male circumcision (WHO 2007)`

Normalization: Disease Incidence Score

- Columns with different units:
  - `Incidence of TB`: per 100,000 people
  - `Diabetes Prevalence`, `Hypertension`: in percentage (
- Method: Min-Max Normalization

$$\text{Normalized Value} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- Combines metrics into a comparable scale.

Handling Screening Coverage Year

- Original values: [2019, 2003, `Not started`, `Unknown`, ...]
- Preprocessing:
  - `Not started` $\rightarrow$ 0
  - `Unknown` $\rightarrow$ NaN
  - Valid years converted to integers

Binning: Male Circumcision

- Raw values: <20, 20–80, >80
- Mapped to:
  - <20 $\rightarrow$ Low
  - 20–80 $\rightarrow$ Medium
  - >80 $\rightarrow$ High

Outlier Detection with IQR

- Method:

$$\text{IQR} = Q_3 - Q_1$$

$$\text{Outliers if: } X < Q_1 - 1.5 \cdot IQR \quad \text{or} \quad X > Q_3 + 1.5 \cdot IQR$$

- Outliers capped or removed to improve stability.

Missing Value Imputation

- Numeric columns: **Median**
- Categorical columns: **Mode**
- Columns with high missing data were dropped.

Dimensionality Reduction using PCA

- Used: `PCA(n_components=0.95)`
- Retains 95% variance with fewer features
- Improves speed and avoids overfitting

Summary

- Cleaned and normalized health indicators
- Converted and mapped categorical values
- Treated outliers with IQR
- Imputed missing values based on type
- Dropped irrelevant/high-missing columns
- Reduced dimensionality with PCA