Data Preprocessing Pipeline for Disease Incidence Analysis

Sainath R.

June 20, 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` $\rightarrow$ 0
  - 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

Model Results – CIN Combined

Model	Train R^2	Test R^2	Train RMSE	Test RMSE
Random Forest (Iter)	0.2329	-0.0962	0.3688	0.4236
XGBoost (Iter)	0.3801	-0.1329	0.3315	0.4306
SVR (Iter)	0.0947	-0.1682	0.4007	0.4373
Ridge (Iter)	0.3886	-0.4333	0.3293	0.4843
XGBoost (Model)	0.9887	-0.5134	0.0448	0.4977

CIN combined parity plots

CIN combined parity plots

SVR Results (with Slack Feature)

Target Variable	Train R^2	Test R^2	Train Rel. RMSE	Test Rel. RMSE
Low CIN (SVR Slack)	-0.0231	-0.1758	0.7006	0.4690
High CIN (SVR Slack)	0.2534	-0.1528	0.2978	0.4833
CIN Combined (SVR Slack)	0.0243	-0.1389	0.4159	0.4317

SVR Results (with Slack Feature)

SVR Results (with Slack Feature)

Low CIN – Model Comparison

Model	Train R^2	Test R^2	Train Rel. RMSE	Test Rel. RMSE
XGBoost (Model Imp)	0.9942	-2.4673	0.0536	0.7208
Ridge (Iterative)	0.3261	-1.4812	0.5686	0.6813

Low CIN – Model Comparison)