Cervical Cancer Data Analysis

Sainath Raja

NIT Raipur

June 6, 2025

Introduction

- Focused on data scraping and preprocessing for Indian states.
- Target: Cervical cancer-related features.
- Also explored global datasets from WHO.

Collected Covariates

- Region wise data
- Total Fertility Rate (2016–2018)
- Cervical cancer by age group
- Diagnosis methods
- City-wise cases (CR, AAR, TR)
- Contraception data
- Doctors per state (2010–2020)
- Schooling (expected/mean)
- Per capita income
- Anaemia data
- Diabetes Hypertension
- HDI
- HIV
- Life expectancy
- Marital age
- Population
- TB data

Model Performance: Low CIN Combined

Model	**Train** R^2	**Test** R^2	**Train Rel. RMSE**	**Test Rel. RMSE**
Ridge	0.1610	-2.3919	0.9160	1.8417
Lasso	0.1587	-1.9439	0.9172	1.7158
Random Forest	0.8525	-1.5410	0.3841	1.5941
SVR (poly)	0.0130	**0.0574**	0.9935	**0.9709**

Best Generalizing Model: SVR with polynomial kernel.

Model Performance: High CIN Combined

Model	**Train R^2**	**Test R^2**	**Train Rel. RMSE**	**Test Rel. RMSE**
Random Forest	0.8294	-0.3623	0.4130	1.1672
XGBoost	1.0000	-1.3537	0.0004	1.5342
Ridge	0.7455	-1.0004	0.5045	1.4143
SVR (sigmoid)	0.1371	**0.0597**	0.9289	**0.9697**

Best Generalizing Model: SVR with sigmoid kernel.

Note on Global Data

- Global data was used for training models on Low CIN, High CIN, and ICC combined prevalence.
- WHO websites were used for population, STIs, and related data.