HPV Prevalence Prediction using SMOGN

Worldwide and Indian States Analysis

Sainath Raja

NIT Raipur

July 4, 2025

Introduction

- Human Papillomavirus (HPV) is a leading cause of cervical cancer.
- Predicting prevalence helps in early intervention and policy-making.
- This project aims to predict HPV prevalence across countries and Indian states.

Dataset and Features

Global Dataset:

- Target Variables: **NCC Combined**, **Low CIN Combined**, **High CIN Combined**
- Features used:
    - Life Expectancy, HIV, HDI, Diabetes, Hypertension, TB
    - Expected Schooling, Doctors per 1000, Population
    - Mean Schooling, Marital Age, Contraception Use
    - Anaemia, Total Fertility Rate, Mortality Rate, Per Capita Income

Worldwide Prediction using SMOGN

- Applied **SMOGN** (Synthetic Minority Oversampling TEchnique for Regression with Gaussian Noise)
- Addressed imbalanced data in regression tasks
- Achieved better predictions for rare prevalence values

Prediction for Indian States

- Collecting Indian-state-level data for the same features
- Scraping prevalence data from published articles
- Will apply the trained SMOGN model for predictions once data is ready

Backup Strategy

- If Indian-state prevalence is not available:
    - Identify countries with similar **HDI**
    - Use their prevalence values as approximations
- This allows for **approximate mapping** to inform policy

Conclusion and Next Steps

- Completed global prevalence prediction using SMOGN
- Currently scraping Indian-state data and collecting prevalence from literature
- Plan to run final prediction model once data collection is done