# Predicting HPV Infection Rates in World wide Using Machine Learning

SAINATH R

June 19, 2025

# Project Overview

- This project aims to predict HPV infection rates across Indian states, where current data is limited.
- It uses machine learning models trained on data from countries with similar development levels.
- The goal is to help estimate cervical cancer risk and improve screening or vaccination strategies.

# Top Covariates Used

- Anemia prevalence among women (15–49 years)
- Physicians per 1,000 people
- Population estimate
- Smoking prevalence (female, 2016)
- Total fertility rate (2017)
- Contraception use (2019)
- HIV prevalence (adults)
- Multiple pregnancies (%)
- Male circumcision (WHO 2007)
- Condom use
- Start year of screening coverage
- Human Development Index (HDI)
- Life expectancy at birth
- Incidence of tuberculosis (per 100k)
- Diabetes, Hypertension prevalence (%)
- HPV Vaccine, STIs

# HPV-Related Prevalence Targets

1. NCC-16-prevalence
2. NCC-18-prevalence
3. Low CIN-16-prevalence
4. Low CIN-18-prevalence
5. High CIN-16-prevalence
6. High CIN-18-prevalence
7. ICC-16-any-prevalence
8. ICC-16-SCC-prevalence
9. ICC-16-ADC-prevalence
10. ICC-18-any-prevalence
11. ICC-18-SCC-prevalence
12. ICC-18-ADC-prevalence

# Data Preprocessing - Dropped Columns

- **Dropped:** Continent, all case counts, aggregate CIN/ICC prevalence columns
- **Objective:** Simplify data to focus on relevant predictors and targets

# Feature Engineering - Disease Incidence Score

- Columns with different units (e.g., TB per 100k vs Diabetes %).
- Normalization used:

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

- TB incidence scaled using population estimate for consistency.

# Handling Categorical and Special Columns

- **Screening Year:**
  - Not started / Unknown $\rightarrow$ 0
  - Year values converted to integers
- **Male Circumcision:**

# Missing Value Imputation

- Numeric columns: **Median Imputation**
- Categorical columns: **Mode Imputation**

# Dimensionality Reduction - PCA

- Used `PCA(n_components=0.95)`
- Retains 95% variance with fewer features

# Train-Test Split and Outliers

- Train: 80% (used for CV and training)
- Test: 20% (final evaluation)
- Outlier Detection included during training

# Model Results – High CIN Combined

| Model | Train $R^2$ | Test $R^2$ | Train RMSE | Test RMSE |
|---|---|---|---|---|
| Random Forest | 0.6057 | -0.2965 | 0.2164 | 0.5126 |
| Ridge (Iterative) | 0.4434 | -0.3384 | 0.2572 | 0.5208 |
| XGBoost (Model Imp) | 1.0000 | -1.3537 | 0.0004 | 1.5342 |
| SVR (Sigmoid) | 0.1371 | 0.0597 | 0.9289 | 0.9697 |

# Model Results – Low CIN Combined

| Model | Train $R^2$ | Test $R^2$ | Train RMSE | Test RMSE |
|---|---|---|---|---|
| Ridge (Iterative) | 0.3261 | -1.4812 | 0.5686 | 0.6813 |
| XGBoost (Model Imp) | 0.9942 | -2.4673 | 0.0536 | 0.7208 |
| Random Forest | 0.8525 | -1.5410 | 0.3841 | 1.5941 |
| SVR (Poly) | 0.0130 | 0.0574 | 0.9935 | 0.9709 |

# 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 |

# Next Steps

- Continue data scraping of Indian states from NCDIR reports and research articles
- Train final models on enriched Indian datasets
- Validate predictions with real HPV burden studies

# Future Work

- Extend the current model to make state-wise HPV prevalence predictions for Indian states using newly scraped data from NCDIR reports and research articles.
- Improve model generalization and predictive power, particularly enhancing $R^2$ scores on the test set.
- Validate predicted prevalence against epidemiological studies or surveys as they become available.
- Explore spatial modeling and temporal trends to better understand region-specific HPV risk patterns.