Case Study: Machine Learing Approach to Predict HPV Prevalence Using Limited Data

Preamble

Cervical cancer remains a significant public health challenge, particularly in low- and middle-income countries, where access to early screening and vaccination is limited. Persistent infection with high-risk Human Papillomavirus (HPV) genotypes—especially HPV-16 and HPV-18—is the primary cause of cervical intraepithelial neoplasia (CIN) and cervical cancer. This case study aims to predict the prevalence of HPV-associated cervical lesions by leveraging machine learning techniques on curated global datasets. Key outcome variables include **NCC_combined**, **Low_CIN_combined**, and **High_CIN_combined**, representing the average burden of lesions caused by both HPV-16 and HPV-18. After data collection and preprocessing, the SMOGN (Synthetic Minority Over-sampling Technique for Regression with Gaussian Noise) method was applied to address data imbalance. Among the tested models—Lasso, Ridge, and XGBoost Regressor—the XGBoost model achieved the best predictive performance. The results demonstrate the potential of data-driven approaches to support both global and regional cervical cancer prevention efforts.

Introduction

Persistent Human Papillomavirus (HPV) infection of the cervix is the leading cause of cervical cancer, which is the third most common cancer in women globally and the second most common in India. Among the 200 HPV genotypes detected; the major oncogenic high risk HPV genotypes are HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66 and 68; of which about 70% of cervical cancer cases have been attributed to HPV16 and HPV18. Studies have reported that globally around 291 million women carry HPV DNA, of whom 32% are infected with HPV16 or HPV18, or both. This also remains a persistent threat to women with chronic HPV infections, increasing their likelihood of developing precancerous lesions that may progress to invasive cervical cancer. In India alone, about 5.0% of women in the general population are estimated to harbor cervical HPV-16/18 infection at a given time, and 83.2% of invasive cervical cancers are attributed to HPV 16 or 18. Moreover, since HPV infections do not present obvious clinical symptoms (with most HPV infections resolving spontaneously), estimates of prevalence and incidence are important to understand the infection burden and etiology in a population.

The world bank data statistics (for the year 2023) show that approximately 471 million Indian women are in the age group of 15 to 64 years. This population is at risk for developing invasive cervical cancer if not screened for/vaccinated (earlier age groups) against HPV infection. Cervical cancer burden data has been actively maintained for all states by organizations like the National Cancer Registries. However, data on HPV prevalence and relevant covariates—such as sexually transmitted infections, hormonal contraceptive use, smoking, early sexual intercourse, multiple sexual partners, high parity, and early pregnancy—remains sparse due to the lack of screening procedures. For

instance, the national average prevalence of cervical cancer screening among women in the age group 30–49 years is 1.97% (1.85–2.09%, 95% C.I), highlighting the need for adequate screening strategies. The lack of comprehensive national screening and vaccination programs in India at present makes it crucial to determine HPV prevalence and identify HPV infection vulnerable population pockets. This can facilitate the timely and appropriate interventions leading to an improved prognosis.

Machine-learning algorithms have shown great promise in predicting the spread and onset of infectious diseases. For example, some studies have used machine learning to forecast the number of cases of a particular disease in a region based on past data and current conditions. The use of machine learning in the prediction of infectious diseases is a promising area of research, with potential applications in public health, epidemiology, and clinical practice

In this data-driven study, we aim to model and predict the prevalence of HPV-associated cervical lesions by focusing on combined indicators that represent the burden of both HPV-16 and HPV-18. Specifically, the study targets the following outcome variables:

- **NCC_combined**: Represents the average prevalence of cervical lesions associated with HPV-16 and HPV-18, combining NCC-16-prevalence and NCC-18-prevalence.
- **Low_CIN_combined**: Combines Low CIN-16-prevalence and Low CIN-18-prevalence, indicating the average prevalence of low-grade cervical intraepithelial neoplasia caused by both genotypes.
- **High_CIN_combined**: Combines High CIN-16-prevalence and High CIN-18-prevalence, representing the average prevalence of high-grade lesions driven by HPV-16 and HPV-18.

While the primary model was developed using global health datasets, we extended its application to Indian states to predict regional HPV lesion prevalence using the same framework. This additional analysis provides insights into the spatial distribution of cervical cancer risk within India and supports state-specific planning for screening and vaccination initiatives. Persistent HPV infection is highly prevalent in women globally and is the causative factor for over 99% of cervical cancer cases. Countries with limited access to effective HPV screening and vaccination programs show disproportionately higher morbidity and mortality. Modeling HPV lesion prevalence can support timely interventions and tailored healthcare strategies, especially in regions with missing or unreliable surveillance data.

This becomes especially relevant in a country like India, where inter-state variation in healthcare access, awareness, and socioeconomic factors can significantly impact HPV-related disease outcomes. Estimating HPV lesion prevalence across Indian states helps address these disparities and prioritize interventions where they are needed most.

Objective

- To clean, preprocess, and unify disparate datasets for consistency and model readiness.

- To engineer relevant features—including demographic, epidemiological, and socioeconomic indicators—to improve model accuracy.
- To experiment with and compare multiple machine learning models (Lasso, Ridge, and XGBoost) to identify the best performer.
- To apply SMOGN (Synthetic Minority Over-sampling Technique for Regression with Gaussian Noise) to address data imbalance and enhance prediction quality.
- To apply the optimized model to both global and Indian datasets for predicting HPV lesion prevalence.
- To visualize final predictions and insights, including:
  - Predicted prevalence for missing countries and Indian states
  - Feature importance analysis
  - Parity plots and residual diagnostics

Methodology

To develop robust and regionally relevant models for predicting HPV-associated lesion prevalence, both global and Indian datasets were collected and curated using a combination of manual extraction, automated PDF scraping, and web research.

Data Sources

- **WHO Population Statistics** (link)
  Country-wise female population data was obtained from WHO databases to standardize prevalence and calculate derived indicators (e.g., incidence per 100,000 females).
  (Link)
- **Scientific Articles**
  WHO data was incomplete (especially for state-wise mortality rates
  (link) or HPV prevalence (link) in India), relevant values were extracted from peer-reviewed research articles and public domain PDFs.

Data Preprocessing

The data preprocessing phase was essential for transforming raw datasets into a clean, consistent, and model-ready format. Below are the key steps followed during this process:

1. **Column Filtering and Data Cleaning**

- Removed irrelevant or redundant columns, such as:
  - Raw case counts (e.g., NCC-16-cases, CIN-18-cases)
  - Intermediate prevalence columns
  - Columns with excessive missing data (e.g., sample size studied)
  - Metadata like "Continent" not relevant for prediction
- Resulted in a lean dataset focused only on meaningful variables

Feature Engineering

1. **Disease Incidence Score**
   - The dataset had three indicators: TB incidence (per 100,000), diabetes prevalence, and hypertension (%).
   - Converted TB incidence to percentage to align units.
   - Applied Min-Max normalization to all three.
   - Averaged the normalized values to create a single `Disease_Incidence_Score`.
   - **Formula:**
     - $\text{TB\_Rate} = \left(\frac{\text{Incidence of TB}}{\text{Population Estimate}}\right) \times 100$
     - $x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$
2. **Screening Coverage Year Conversion**
   - Original values included both years and labels like "Not started" or "Unknown".
   - Converted "Not started" and "Unknown" to 0.
   - Retained valid years as integers.
   - Created a clean numeric variable representing the start year of screening.
3. **Male Circumcision Binning**
   - Original entries: "¡20", "20–80", "¿80"
   - Mapped these to ordinal categories:
     - less than 20 $\rightarrow$ **Low**
     - 20–80 $\rightarrow$ **Medium**
     - greater than 80 $\rightarrow$ **High**
   - Improved interpretability and enabled categorical encoding.

- **STI Burden Score**

  - Dataset included STI rates: Syphilis, Chlamydia, and Gonorrhea.
  - Normalized each individually using Min-Max scaling.
  - Averaged them to create a single `STI_Score`.
  - Captured general sexual health burden in each region.
  - **Equation:**
    * $x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$
    * STI_Score $= \frac{1}{n} \sum x_{i_{\text{norm}}}, \quad$ where $n = 3$

Target Variable Construction

- Created three combined targets:
  - `NCC_combined`: Average of NCC-16 and NCC-18 prevalence
  - `Low_CIN_combined`: Average of low-grade CIN due to HPV-16 and HPV-18
  - `High_CIN_combined`: Average of high-grade CIN due to both genotypes
- These targets reduced variability and ensured balanced prediction goals.

Handling Missing Values

- **Numerical columns:**
  - Filled using median to reduce sensitivity to outliers
- **Categorical columns:**
  - Filled using mode (most frequent value)
- **Target columns:**
  - Rows with missing target values were removed
- **High-nullity columns:**
  - Dropped if missing values exceeded acceptable thresholds

Normalization and Encoding

- Applied Min-Max scaling where variables had different ranges
- Categorical variables (e.g., circumcision category) prepared for one-hot encoding
- Ensured compatibility with both tree-based (XGBoost) and linear models (Lasso, Ridge)

Model Building

This section details the machine learning models applied, the issues faced during initial model experimentation, and the final approach that significantly improved model performance. A range of regression models were used to predict the three target variables: `Low_CIN_combined`, `High_CIN_combined`, and `NCC_combined`. The model building process included extensive experimentation with various regression algorithms, imputation techniques, and hyperparameter tuning strategies.

Initial Model Experimentation

To begin, traditional regression models were trained on the original dataset, without any imbalance handling techniques. The following models were explored:

- Random Forest Regressor
- Ridge Regression (with Iterative Imputation)
- XGBoost Regressor
- Support Vector Regressor (SVR)

To overcome the imbalance problem, the SMOGN (Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise) (link) algorithm was adopted. This technique generates synthetic samples in underrepresented regions of the target variable distribution, thereby improving the model's ability to generalize in sparse data zones.

The decision to use SMOGN was informed by its demonstrated effectiveness in small, imbalanced regression problems, as supported by findings published in the *International Journal of Medical Informatics*. In these studies, SMOGN outperformed traditional resampling techniques by not only addressing imbalance but also preserving the original data structure, reducing noise sensitivity, and enhancing prediction quality.

Given the skewed distribution of HPV lesion prevalence—especially in low-income countries and underreported Indian states—SMOGN was particularly well-suited to ensure that the model remained sensitive to rare but clinically significant cases.

MODEL WORKFLOW

```
+-- 1. Data Loading
|   +-- Read Excel file

+-- 2. Data Cleaning
|   +-- Drop unnecessary columns
|   +-- Handle special cases

+-- 3. Feature Engineering
|   +-- TB_Incidence_Percent = (Incidence of TB / Population) x 100
|   +-- Disease_Incidence_Score = Avg(TB, Hypertension, Diabetes)
|   +-- STI_Score = Avg(Syphilis, Chlamydia, Gonorrhea normalized rates)
|   +-- Male Circumcision Category (Low / Medium / High)

+-- 4. Drop Extra Columns
|   +-- Drop all intermediate and unused prevalence/case columns

+-- 5. Feature & Target Split
|   +-- X = Features
|   +-- y = Target

+-- 6. Encoding & Imputation
|   +-- Drop high-cardinality fields: Country, Economy
|   +-- One-hot encode categorical fields
|   +-- Impute numeric & categorical values (Median / Most Frequent)

+-- 7. Feature Selection
|   +-- Use LassoCV to rank features
|   +-- Select non-zero coefficient features using SelectFromModel

+-- 8. SMOGN Oversampling
|   +-- Applied on full dataset (X + y)
|   +-- Synthesizes minority samples using Gaussian noise

+-- 9. Train-Test Split
|   +-- 80% Train, 20% Test using random_state=42

+-- 10. Model Pipeline
```

```
|   +-- Preprocessor (Scaling + OneHotEncoding)
|   +-- PCA (Retain 95% variance)
|   +-- XGBoostRegressor with tuned hyperparameters

+-- 11. Model Training
|   +-- Fit final pipeline on training set

+-- 12. Model Evaluation
|   +-- Train R², Test R²
|   +-- Train Rel RMSE, Test Rel RMSE

+-- 13. Integrity Checks
|   +-- No data leakage | Target used only as y
|   +-- Pipeline ensures same transformations on train & test
|   +-- SMOGN applied before split (justified for experimentation)
```

List of Covariates Used in the Global Model

S. No.	Covariate Name	Description
1	Anemia prevalence among non-pregnant women (%)	Fraction of non-pregnant women ages 15–49 with anemia
2	Physicians per 1,000 people	Density of physicians in the population
3	Anemia prevalence among pregnant women (%)	Fraction of pregnant women ages 15–49 with anemia
4	Mean targeted age	Average or representative age group used for studies/screening
5	Population estimate	Estimated population size of the targeted age group
6	Smoking Prevalence (females, 2016)	Percentage of currently smoking females
7	Total Fertility rate (2017)	Average number of births per woman
8	Contraception use (2019)	Use of oral, injectable, or implant contraceptives among females (%)
9	HIV Prevalence (in adults)	HIV prevalence rate (%) in the adult population
10	Mean marital age	Average age at which individuals (usually females) marry
11	Male circumcision (WHO 2007)	Percentage or category (Low/Medium/High) of men circumcised
12	Condom Use	Proportion (%) of sexually active population using condoms
13	Start of Screening coverage (year)	Year in which cervical screening programs began
14	HPV vaccination introduction	Year of HPV vaccine introduction into national immunization programs
15	Age-adjusted incidence (standardized rates)	Standardized incidence rates for cervical cancer (per 100,000 or %)
16	Number of deaths (all ages, 2021)	Count of cervical cancer-related deaths in 2021
17	Mortality rates (age standardized)	Age-standardized mortality rates for cervical cancer
18	Human Development Index (HDI)	Composite index (0–1) of life expectancy, education, and income
19	Life expectancy at birth	Estimated life expectancy at birth
20	Expected years of schooling	Projected total years of formal education
21	Mean years of schooling	Average years of schooling for adults aged 25+
22	Gross national income (GNI) per capita	Per-person national income
23	Incidence of TB	Tuberculosis incidence rate (per 100,000 or %)
24	Diabetes Prevalence	Prevalence of diabetes in the adult population (%)
25	HPV Vaccine coverage	Proportion (%) of targeted population vaccinated against HPV
26	Hypertension	Hypertension prevalence (%)

S. No.	Covariate Name	Description
27	Coverage ever screened (women 30–49 years) (%)	Women aged 30–49 who have ever undergone cervical screening
28	Coverage in last 5 years (women 30–49 years) (%)	Women 30–49 screened within last 5 years
29	Coverage in last 3 years (women 30–49 years) (%)	Women 30–49 screened within last 3 years
30	Coverage ever screened (women 25–65 years) (%)	Women 25–65 who have ever undergone screening
31	Coverage in last 5 years (women 25–65 years) (%)	Women 25–65 screened within last 5 years
32	Coverage in last 3 years (women 25–65 years) (%)	Women 25–65 screened within last 3 years
33	STI Female	Prevalence of sexually transmitted infections among females
34	STI combined	Combined prevalence of STIs (both sexes) (% or rate)
35	STI_SyphilisRate	Reported incidence rate of syphilis per 100,000 population in 2021
36	STI_ChlamydialInfectionRate	Reported incidence rate of chlamydia per 100,000 population in 2021
37	STDRates_Gonococcal	Reported incidence rate of gonorrhea per 100,000 population in 2021

Observation Counts for Target Variables

S. No.	Target Variable	Number of Observations
1	NCC-16-prevalence	85
2	NCC-18-prevalence	80
3	Low CIN-16-prevalence	60
4	Low CIN-18-prevalence	52
5	High CIN-16-prevalence	66
6	High CIN-18-prevalence	62