HPV Prevalence Prediction from Missing and Limited Data

Progress Made till June 20th, 2025

Step 1: Data Acquisition

Data for 40 different covariates (features) was scraped off, cleaned, and aggregated from https://hpvcentre.net/ (links to other sources) and literature across various countries (observations).

Step 2: Data Sets Considered

Target Variables	Observations	Features
NCC combined (HPV16 & 18)	99	40
Low CIN combined (HPV16 & 18)	63	40
High CIN combined (HPV16 & 18)	71	40
ICC combined (HPV16 & 18)	90	40

Step 3: Model Training Criteria

15% gold (hold-out) set, 85% training set.

20-fold cross-validation on the training set (Hyperparameter tuning).

Model Building & Data Preparation

Data Preprocessing and Feature Engineering

Step 4: Model Building

Models Considered: LinearRegression, Ridge, Lasso, Random Forest, GradientBoosting, SVR, PCR, XGBOOST.

Best Models Selected for Targets (based on Gold R2):

Target Variables	Model	Gold R^2
NCC combined (HPV16 & 18)	XGBOOST	0.5281
Low CIN combined (HPV16 & 18)	XGBOOST	-0.1840
High CIN combined (HPV16 & 18)	XGBOOST	0.0307
ICC combined (HPV16 & 18)	SVR	-0.2949

Data Preprocessing & Feature Engineering

Dropped Columns:

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

Feature Engineering - Disease Incidence Score:

- Handled columns with different units (e.g., TB per 100k vs Diabetes %).
- **Normalization Used:**

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

- 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:** `<20` $\rightarrow$ LOW; `20 - 80` $\rightarrow$ MEDIUM; `>80` $\rightarrow$ HIGH.

Missing Value Imputation:

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

Dimensionality Reduction - PCA:

- Used `PCA(n_components=0.95)` to retain 95% variance with fewer features.

Future Work & Indian State-wise Prediction

Future Work

Resolving the negative scores of target variables.

Predicting the prevalence of HPV in NCC, CINs, and Invasive Cervical Cancer for Indian states using imputation and existing data in literature.

Extension to Indian States

Objective: Extend the current model to make state-wise HPV prevalence predictions for Indian states.

Approach: Utilizing newly scraped data from NCDIR reports and research articles. This includes data for top cities in India, which will be integrated to enhance the model's predictive capabilities for a more granular, state-level analysis.