

TRIGGER WARNING!

Violence Against Women

Using multi-level datasets and statistical modeling to understand violence risk factors and provide country-specific policy insights.



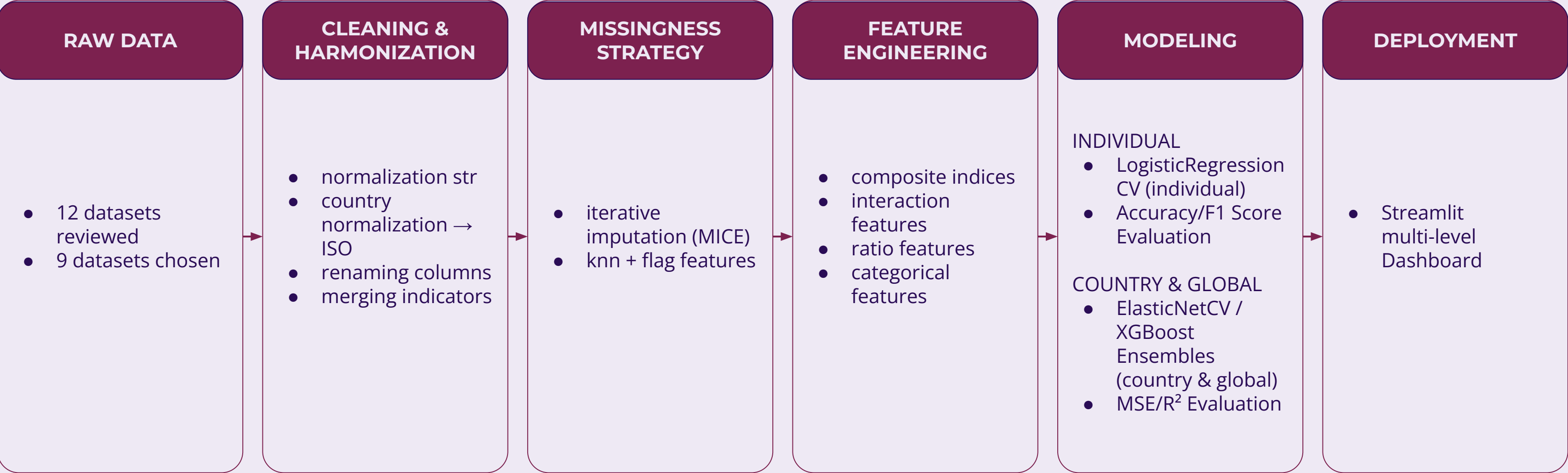
The UN describes violence against women and girls (VAWG) as:

“One of the most widespread, persistent, and devastating human rights violations in our world today. It remains largely unreported due to the impunity, silence, stigma, and shame surrounding it.”

	GOAL		
		<p>To build an evidence-based, multi-level statistical framework that identifies the structural, social, and individual drivers of Violence Against Women & Girls (VAWG) and gain actionable insight to work against them.</p> <p>... <i>“behind every number is a life.”</i></p>	
	Data-based analysis that shed light on the structural conditions that place women at risk.		3

	CONTENT	
		<div><div>DATASETS<ul style="list-style-type: none">● Individual-Level (325 obs.)● Country-Level Violence (50 countries)● Global Development (195 countries)</div><div>METHODS<ul style="list-style-type: none">● Cleaning & Harmonizing● Feature Engineering● Predictive Modeling</div><div>INSIGHTS<ul style="list-style-type: none">● Key Predictors● High-Risk Clusters● Patterns</div><div>APPLICATION<ul style="list-style-type: none">● Streamlit Dashboard</div></div>
	Data-based analysis that shed light on the structural conditions that place women at risk.	4

WORKFLOW



DATASETS

~570 total data entities across three levels

~60–70 cleaned and engineered features

~20 composite/interacted/binned features created through the pipeline

<u>Domestic Violence Against Women</u> age, education, employment, income, marital status, violence
<u>Violence Against Women & Girls</u> country, gender, marital status, education, age, employment, residence, reason, %violence
<u>Legal Frameworks</u> country, % implementation legal framework equality women in marriage
<u>Gender Inequality Index</u> country, GII, human development, maternal mortality, adolescent birth, seats parliament, f/m sec education, f/m labour force
<u>Most Dangerous Countries for Women</u> country, WPSI, WDI, intentional homicide, NPV, IPV, legal discrimination, GGG, GII
<u>The Human Freedom Index</u> country, human freedom index, economic freedom index, personal freedom index
<u>Violence: Intimate Partner Violence (WHO)</u> country, %intimate partner violence
<u>Global Unemployment Data</u> country, sex, age_group, rate
<u>GNI (World Bank)</u> country, GNI (USD)

LEVEL	SIZE	FEATURES
Individual “Why Her?”	n = 325 women	<ul style="list-style-type: none">age, employment, educationrelationship factorsprevious exposure to violencehousehold economic stress Target: violence (binary)
Country-Level Violence “Which societal attitudes predict higher IPV rates?”	n = 50 countries	<ul style="list-style-type: none">national violence prevalenceyouth unemploymentgender gapsparliamentary representationlegal discriminationmaternal health indicatorsfreedom indices Target: violence (continuous)
Global-Level Violence “Which country-level interventions reduce violence?”	n = 195 countries ~50 indicators	<ul style="list-style-type: none">economic development (GNI, HDI)gender inequality indicesfreedom and governance measuresyouth vulnerability indicatorsmaternal health indicators Target: violence (continuous)

METHODS: Key Technical Components

1. Missing Data Strategy

- *IterativeImputer* for numeric variables
- *Median/KNN fallback* when needed
- Missingness flags ➡ improves model performance

2. Composite Features (PCA + thematic indices)

- **Economic Development Index (PCA)**
- **Gender Inequality Composite (PCA)**
- **Youth Vulnerability Index**
- **Maternal Health Risk Index**

Many global indicators are correlated ➡ PCA reduces redundancy

3. Interaction Features

E.g.,

- **economic_development × gender_inequality**
- **freedom × inequality**
- **political power × economic power**

These capture real-world combined effects.

4. Ensemble Model

Why ensembles:

- Handle non-linearities
- Perform well on country-level small-N datasets
- Avoid overfitting with cross-validation

KEY INSIGHTS: EDA & Modeling

INDIVIDUAL-LEVEL INSIGHTS

Strongest predictors of experiencing violence:

- Prior exposure to violence
- Age (younger women slightly higher risk)
- Economic stress (no- to low-income at higher risk)
- Relationship factors (married at higher risk)

COUNTRY-LEVEL VIOLENCE INSIGHTS

Structural drivers:

- Gender inequality (education gap, labor force gap)
- Youth unemployment
- Political representation (women’s seats in parliament)
- Economic precarity

Clustering revealed:

- High inequality + high youth unemployment cluster = highest rates

GLOBAL INSIGHTS

Top predictors of IPV prevalence (global model):

- Economic development index (↓ violence with ↑ development)
- Gender inequality composite
- Maternal health risk
- Freedom index
- Youth vulnerability index

These appear repeatedly across all models = **robust drivers**.

MODEL PERFORMANCE

- Individual Model:
 - LinearRegressionCV with ElasticNet Penalty
 - F1: 0.94, Accuracy: 0.94
- Violence Model:
 - ElasticNetCV
 - R² Train/Test 1.00 ➡ small n, model remembers
- Global Model:
 - ElasticNetCV + XGBoost Ensemble
 - R² Train: 0.487; R² Test 0.264

VAWG Dashboard

LOCAL HOST

KEY CHALLENGES & LIMITATIONS

Challenge 1 — Heterogeneous data sources

- Different structures
- Missing values, inconsistent country names
- Different units/scales ➡ Required standardization, harmonization, and sophisticated imputation.

Challenge 2 — Multi-level nature of VAWG

VAWG is influenced simultaneously by:

- individual circumstances
- relationship/household dynamics
- community and national structures
- gender inequalities embedded in society

No *single* dataset captures all this.

Splitting into **three levels** was not just practical, it was conceptually necessary.

Challenge 3 — Small sample sizes at macro levels

- Country-level violence dataset has only **50 countries**
- High risk of overfitting
 - ➡ Required feature engineering, dimensionality reduction (PCA), and ensemble models.

Challenge 4 — Missingness as information

Missing data was *not random*: lower-income or conflict-affected countries had systematically missing indicators.

- missingness flags
- iterative imputation
- careful checks for imputation bias

Challenge 5 — Combining quantitative modeling with a human topic

Predicting violence carries ethical weight.

You had to balance:

- scientific rigor
- interpretability
- sensitivity to the human implications

CONCLUSION

- Integration of multi-level data to model VAWG risk is possible.
- Composite + interaction features improve performance and interpretability.
- Ensemble models capture structural drivers better than linear models alone.
- A dashboard can make insights accessible to practitioners.

***Violence Against Women & Girls is not inevitable:
It is shaped by structures, economies, inequalities, and choices made at every level of society.***

THANK YOU!



*Never,
under any
circumstance,
a woman
"deserves to
be beaten".*