Violence against women Data analysis using PCA, Logistic Regression and Decision Trees

Gherasim Delia-Cătălina
Applied Computational Intelligence
Group 246-1

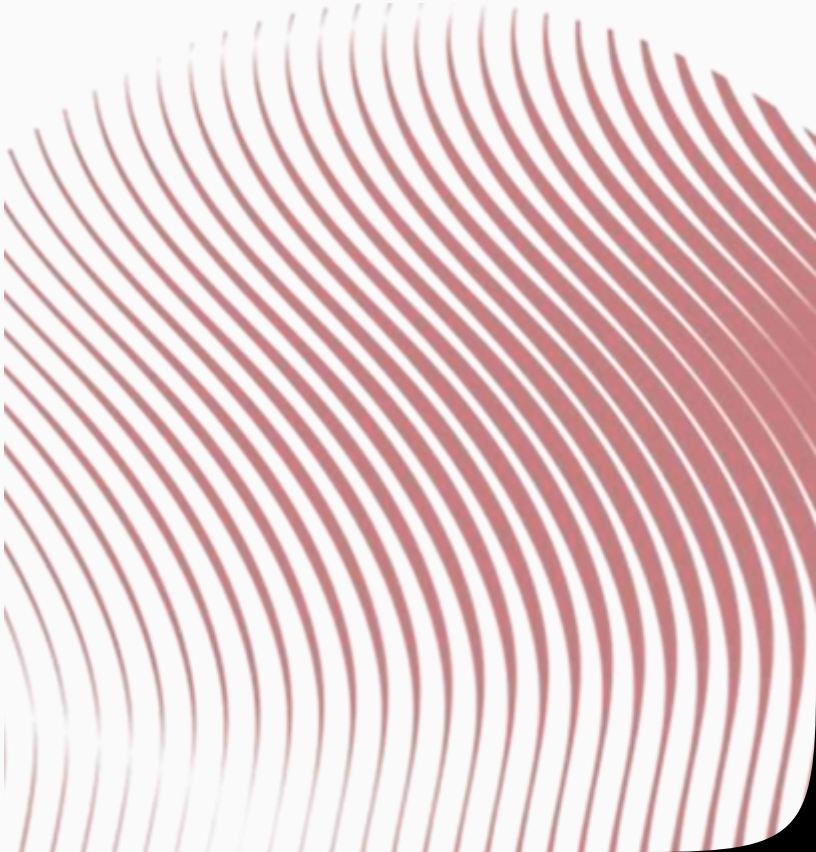


Table of contents

01 Introduction

Intimate partner violence (IPV) remains a critical global health and human rights challenge.

Understanding the justification of violence is important for developing effective preventive measures. The acceptance of violence allows harmful norms to persist, serving as a powerful barrier to policy implementation and change.

Related work

Hacıaliefendioğlu A, Y. S. (2020). Co-occurrence patterns of intimate partner violence. *In BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, 79-90.

Anasel, M. G. (2024, Nov 25). A Logit Regression Approach to Identifying Predictors of Intimate Partner Violence in Tanzania. *East African Journal of Applied Health Monitoring and Evaluation 7, no. 2*.

Coll CV, S. T. (2021). Identifying the women most vulnerable to intimate partner violence: A decision tree analysis from 48 low and middle-income countries. EClinicalMedicine 42.

03 Theoretical aspects

1. Principal component analysis

Singular value decomposition: **Scores** (positions of observations in the new PCA space), **Loadings** (how strongly each original variable contributes to each component), and **Singular values** (how much variance each component explains).

2. Logistic regression

Logistic regression is used when the outcome is binary. The logit is the natural logarithm of the odds of the event happening

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k$$

3. Decision trees

A decision tree predicts an outcome by repeatedly splitting the data into smaller groups.

Each path from the root to a leaf forms an **if-then rule**.

“Violence against women & girls” dataset

The percentage of people in the survey group who agree with the statements:
A husband is justified in hitting or beating his wife…

- if she burns the food
- if she argues with him
- if she goes out without telling him
- If she neglects the children
- if she refuses to have sex with him
- for at least one specific reason

```
RangeIndex: 12600 entries, 0 to 12599
Data columns (total 8 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   RecordID                12600 non-null  int64
 1   Country                 12600 non-null  object
 2   Gender                  12600 non-null  object
 3   Demographics Question   12600 non-null  object
 4   Demographics Response   12600 non-null  object
 5   Question                12600 non-null  object
 6   Survey Year             12600 non-null  object
 7   Value                   11187 non-null  float64
```

	RecordID	Country	Gender	Demographics Question	Demographics Response	Question	Survey Year	Value
0	1	Afghanistan	F	Marital status	Never married	... if she burns the food	01/01/2015	NaN
1	1	Afghanistan	F	Education	Higher	... if she burns the food	01/01/2015	10.1
2	1	Afghanistan	F	Education	Secondary	... if she burns the food	01/01/2015	13.7
3	1	Afghanistan	F	Education	Primary	... if she burns the food	01/01/2015	13.8
4	1	Afghanistan	F	Marital status	Widowed, divorced, separated	... if she burns the food	01/01/2015	13.8

05 Preprocessing

Date handling

Merging question responses

Pivoted to create one column per survey question.

Handling missing data

Replaced with zeros

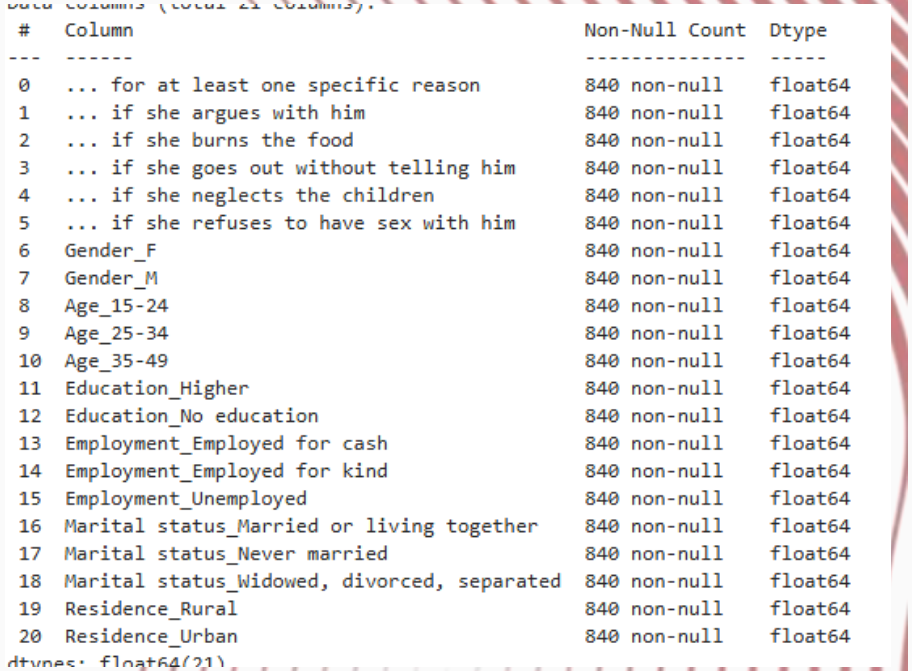
Reshaping demographic information

Reshaped using a pivot table so that each demographic question formed its own column.

Encoding categorical variables

Country, age, education, employment, marital status and residence

Removing non-numeric identifiers



#	Column	Non-Null Count	Dtype
0	... for at least one specific reason	840 non-null	float64
1	... if she argues with him	840 non-null	float64
2	... if she burns the food	840 non-null	float64
3	... if she goes out without telling him	840 non-null	float64
4	... if she neglects the children	840 non-null	float64
5	... if she refuses to have sex with him	840 non-null	float64
6	Gender_F	840 non-null	float64
7	Gender_M	840 non-null	float64
8	Age_15-24	840 non-null	float64
9	Age_25-34	840 non-null	float64
10	Age_35-49	840 non-null	float64
11	Education_Higher	840 non-null	float64
12	Education_No education	840 non-null	float64
13	Employment_Employed for cash	840 non-null	float64
14	Employment_Employed for kind	840 non-null	float64
15	Employment_Unemployed	840 non-null	float64
16	Marital status_Married or living together	840 non-null	float64
17	Marital status_Never married	840 non-null	float64
18	Marital status_Widowed, divorced, separated	840 non-null	float64
19	Residence_Rural	840 non-null	float64
20	Residence_Urban	840 non-null	float64

dtypes: float64(21)

Principle component analysis

Label encoding and one-hot encoding
StandardScaler

Logistic regression

One-hot encoding
Standard scaler

Decision trees

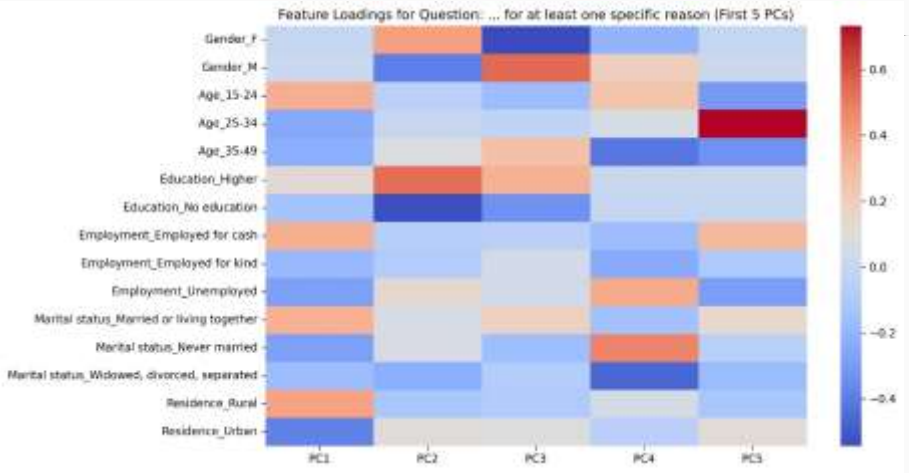
Synthetic Minority Oversampling Technique for Nominal and Categorical

06 Data analysis

1. Principle component analysis
2. Logistic regression
3. Decision trees

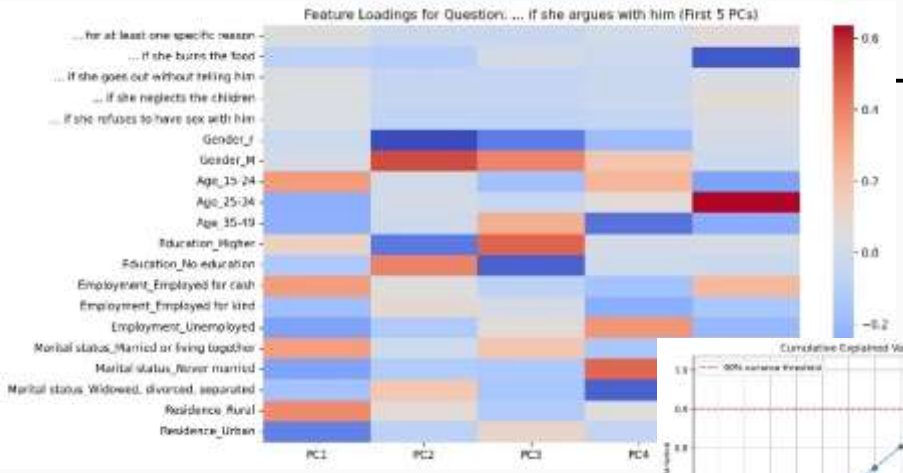
PCA

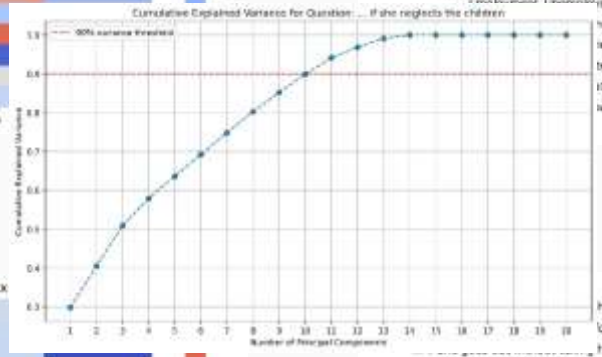
```
Top 10 features for PC1:
Residence                                0.529956
Age                                      0.468044
Employment                               0.454090
Marital status                           0.435216
Country                                  0.189586
Education                                0.178916
... if she burns the food                0.128922
... for at least one specific reason     0.068764
... if she neglects the children         0.067992
... if she goes out without telling him  0.061080
Name: PC1, dtype: float64

Top 10 features for PC2:
... if she burns the food                0.690587
Country                                  0.565352
Education                                0.327381
... for at least one specific reason     0.191827
Gender                                   0.138846
... if she goes out without telling him  0.112881
Marital status                           0.083854
... if she neglects the children         0.082210
Age                                      0.076118
... if she argues with him               0.055759
Name: PC2, dtype: float64
```



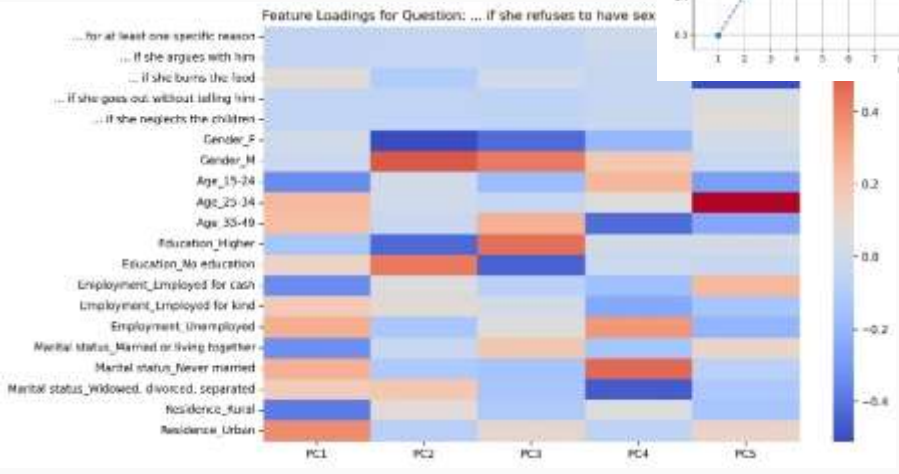
PCA on all features using label encoding	9
PCA on all features using one-hot encoding	11
PCA performed separately for each survey question	10
PCA per question while excluding all other survey questions	10

Number of components needed to obtain 90% variance


Feature Loadings for Question: ... if she argues with him (First 5 PCs)
... for at least one specific reason
... if she burns the food
... if she goes out without telling him
... if she neglects the children
... if she refuses to have sex with him
Gender_F
Gender_M
Age_15-24
Age_25-34
Age_35-49
Education_Higher
Education_No education
Employment_Employed for cash
Employment_Employed for kind
Employment_Unemployed
Marital status_Married or living together
Marital status_Never married
Marital status_Widowed, divorced, separated
Residence_Rural
Residence_Urban
PC1
PC2
PC3
PC4
0.6
0.4
0.2
0.0
−0.2


Cumulative Explained Variance for Question: ... if she neglects the children
90% variance threshold
Number of Principal Components


Feature Loadings for Question: ... if she goes out without telling him (First 5 PCs)
... for at least one specific reason
... if she argues with him
... if she burns the food
... if she neglects the children
... if she refuses to have sex with him
Gender_F
Gender_M
Age_15-24
Age_25-34
Age_35-49
Education_Higher
Education_No education
Employment_Employed for cash
Employment_Employed for kind
PC1
PC2
PC3
PC4
PC5
0.6
0.4
0.2
0.0
−0.2
−0.4


Feature Loadings for Question: ... if she refuses to have sex
... for at least one specific reason
... if she argues with him
... if she burns the food
... if she goes out without telling him
... if she neglects the children
Gender_F
Gender_M
Age_15-24
Age_25-34
Age_35-49
Education_Higher
Education_No education
Employment_Employed for cash
Employment_Employed for kind
Employment_Unemployed
Marital status_Married or living together
Marital status_Never married
Marital status_Widowed, divorced, separated
Residence_Rural
Residence_Urban
PC1
PC2
PC3
PC4
PC5
0.4
0.2
0.0
−0.2
−0.4


Feature Loadings for Question: ... for at least one specific reason (First 5 PCs)
... if she argues with him
... if she burns the food
... if she goes out without telling him
... if she neglects the children
... if she refuses to have sex with him
Gender_F
Gender_M
Age_15-24
Age_25-34
Age_35-49
Education_Higher
Education_No education
Employment_Employed for cash
Employment_Employed for kind
Employment_Unemployed
Marital status_Married or living together
Marital status_Never married
Marital status_Widowed, divorced, separated
Residence_Rural
Residence_Urban
PC1
PC2
PC3
PC4
PC5
0.6
0.4
0.2
0.0
−0.2
−0.4

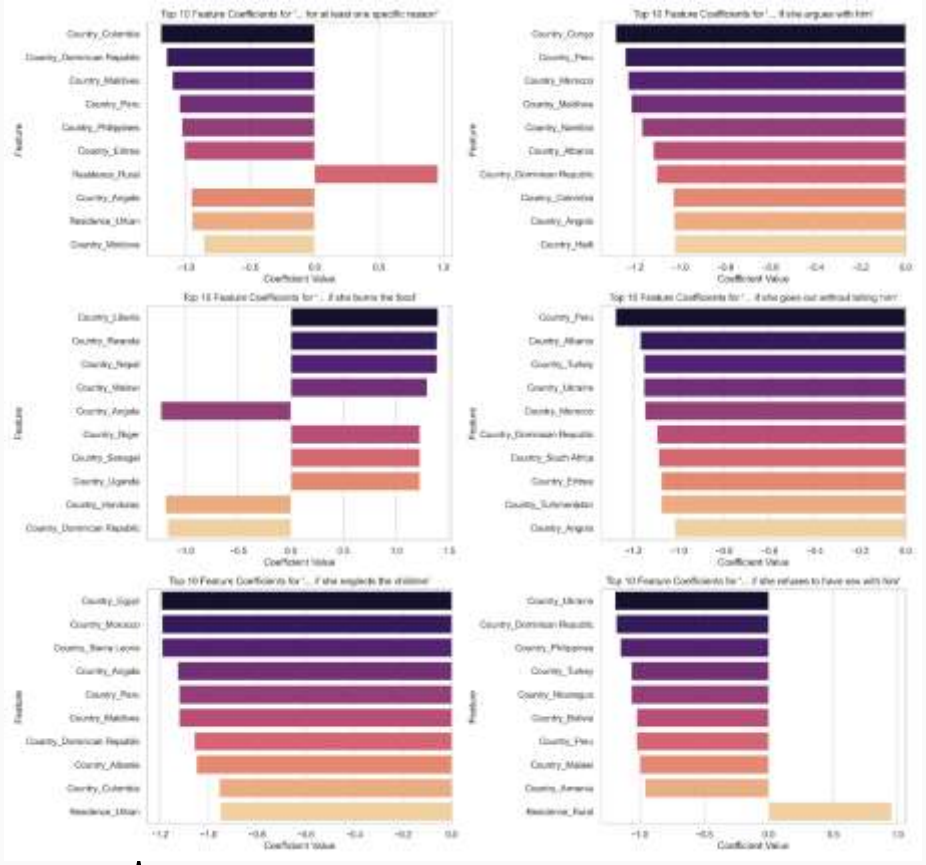
Logistic regression

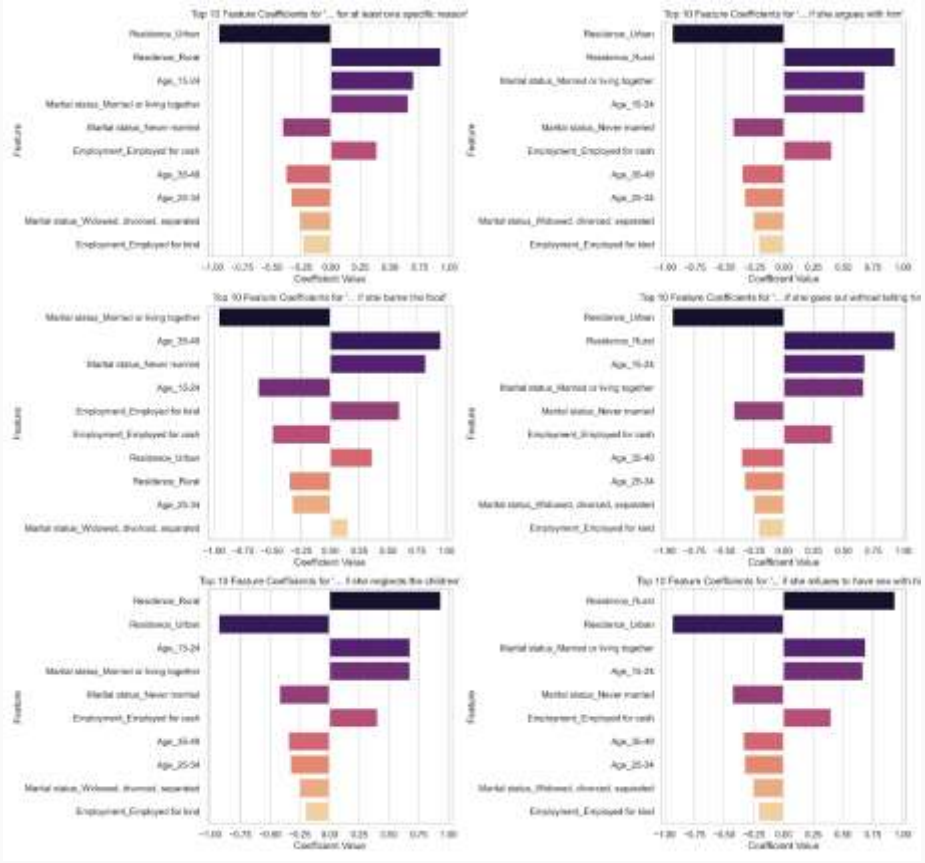
Binary target based on mean

Thresholds based on the mean of the response “yes” percentage for each question were used to define binary outcomes. This preserved the natural distribution of responses. Because many questions had heavily skewed values, the resulting class balance was uneven, limiting predictive performance.

Evenly balanced binary target

This approach aimed to balance classes at approximately 50/50 by adaptively determining thresholds based on sorted response values. This improved the representation of minority responses, but it introduced artificial splits that did not reflect the natural distribution, affecting reliability..


Top 10 Feature Coefficients for '... for at least one specific reason'
Country_Colombia
Country_Dominican Republic
Country_Maldives
Country_Peru
Country_Philippines
Country_Eritrea
Residence_Rural
Country_Angola
Residence_Urban
Country_Moldova
Feature
Coefficient Value
Top 10 Feature Coefficients for '... if she argues with him'
Country_Congo
Country_Peru
Country_Morocco
Country_Maldives
Country_Namibia
Country_Albania
Country_Dominican Republic
Country_Colombia
Country_Angola
Country_Haiti
Feature
Coefficient Value
Top 10 Feature Coefficients for '... if she burns the food'
Country_Liberia
Country_Rwanda
Country_Nepal
Country_Malawi
Country_Angola
Country_Niger
Country_Senegal
Country_Uganda
Country_Honduras
Country_Dominican Republic
Feature
Coefficient Value
Top 10 Feature Coefficients for '... if she goes out without telling him'
Country_Peru
Country_Albania
Country_Turkey
Country_Ukraine
Country_Morocco
Country_Dominican Republic
Country_South Africa
Country_Eritrea
Country_Turkmenistan
Country_Angola
Feature
Coefficient Value
Top 10 Feature Coefficients for '... if she neglects the children'
Country_Egypt
Country_Morocco
Country_Sierra Leone
Country_Angola
Country_Peru
Country_Maldives
Country_Dominican Republic
Country_Albania
Country_Colombia
Residence_Urban
Feature
Coefficient Value
Top 10 Feature Coefficients for '... if she refuses to have sex with him'
Country_Ukraine
Country_Dominican Republic
Country_Philippines
Country_Turkey
Country_Nicaragua
Country_Bolivia
Country_Peru
Country_Malawi
Country_Armenia
Residence_Rural
Feature
Coefficient Value


Top 10 Feature Coefficients for '... for at least one specific reason'
Residence_Urban
Residence_Rural
Age_15-24
Marital status_Married or living together
Marital status_Never married
Employment_Employed for cash
Age_35-49
Age_25-34
Marital status_Widowed, divorced, separated
Employment_Employed for food
Feature
Coefficient Value
Top 10 Feature Coefficients for '... if she argues with him'
Residence_Urban
Residence_Rural
Marital status_Married or living together
Age_15-24
Marital status_Never married
Employment_Employed for cash
Age_35-49
Age_25-34
Marital status_Widowed, divorced, separated
Employment_Employed for food
Feature
Coefficient Value
Top 10 Feature Coefficients for '... if she burns the food'
Marital status_Married or living together
Age_35-49
Marital status_Never married
Age_15-24
Employment_Employed for food
Employment_Employed for cash
Residence_Urban
Residence_Rural
Age_25-34
Marital status_Widowed, divorced, separated
Feature
Coefficient Value
Top 10 Feature Coefficients for '... if she goes out without telling him'
Residence_Urban
Residence_Rural
Age_15-24
Marital status_Married or living together
Marital status_Never married
Employment_Employed for cash
Age_35-49
Age_25-34
Marital status_Widowed, divorced, separated
Employment_Employed for food
Feature
Coefficient Value
Top 10 Feature Coefficients for '... if she neglects the children'
Residence_Rural
Residence_Urban
Age_15-24
Marital status_Married or living together
Marital status_Never married
Employment_Employed for cash
Age_35-49
Age_25-34
Marital status_Widowed, divorced, separated
Employment_Employed for food
Feature
Coefficient Value
Top 10 Feature Coefficients for '... if she refuses to have sex with him'
Residence_Rural
Residence_Urban
Marital status_Married or living together
Age_15-24
Marital status_Never married
Employment_Employed for cash
Age_35-49
Age_25-34
Marital status_Widowed, divorced, separated
Employment_Employed for food
Feature
Coefficient Value

Classification metrics for each question:

	question	accuracy	precision	recall	f1_score	roc_auc
0	... for at least one specific reason	0.495238	0.025641	0.062500	0.036364	0.301703
1	... if she argues with him	0.428571	0.099099	0.354839	0.154930	0.362588
2	... if she burns the food	0.800000	0.341463	0.482759	0.400000	0.743284
3	... if she goes out without telling him	0.428571	0.041667	0.125000	0.062500	0.278968
4	... if she neglects the children	0.404762	0.039604	0.125000	0.060150	0.262904
5	... if she refuses to have sex with him	0.447619	0.052632	0.161290	0.079365	0.325284

Evenly balanced with country

Classification metrics for each question:

	question	accuracy	precision	recall	f1_score	roc_auc
0	... for at least one specific reason	0.228571	0.164948	1.000000	0.283186	0.428195
1	... if she argues with him	0.514286	0.161905	0.548387	0.250000	0.548567
2	... if she burns the food	0.909524	0.812500	0.448276	0.577778	0.689084
3	... if she goes out without telling him	0.228571	0.164948	1.000000	0.283186	0.428195
4	... if she neglects the children	0.228571	0.164948	1.000000	0.283186	0.433813
5	... if she refuses to have sex with him	0.547619	0.180000	0.580645	0.274809	0.580014

Evenly balanced without country

Based on mean
Without country

Classification metrics for each question:

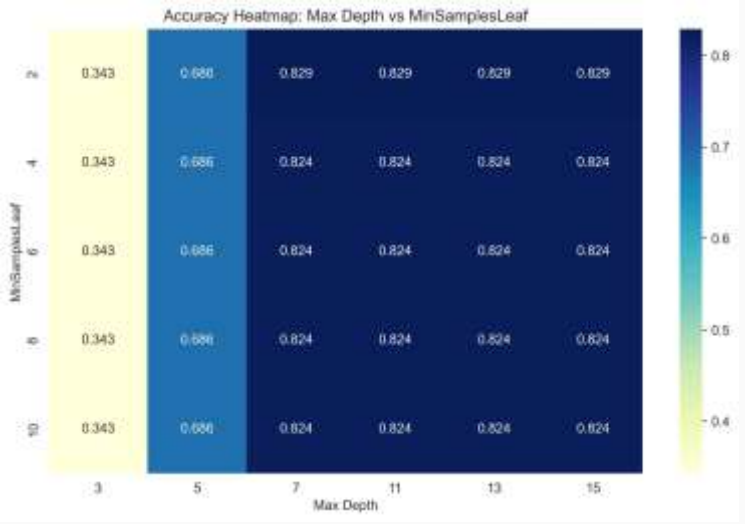
	question	accuracy	precision	recall
0	... for at least one specific reason	0.528571	0.144330	0.466667
1	... if she argues with him	0.542857	0.145833	0.500000
2	... if she burns the food	0.895238	0.538462	0.304348
3	... if she goes out without telling him	0.214286	0.149485	1.000000
4	... if she neglects the children	0.566667	0.164835	0.500000
5	... if she refuses to have sex with him	0.504762	0.111111	0.407407

Based on mean
With country

Classification metrics for each question:

	question	accuracy	precision	recall
0	... for at least one specific reason	0.476190	0.074468	0.233333
1	... if she argues with him	0.423810	0.065421	0.250000
2	... if she burns the food	0.747619	0.187500	0.391304
3	... if she goes out without telling him	0.428571	0.030928	0.103448
4	... if she neglects the children	0.442857	0.051546	0.166667
5	... if she refuses to have sex with him	0.519048	0.097826	0.333333

Decision trees



Without SMOTE

The accuracy was 0.2047 for all depths (3,5,7,10,15)
Severe class imbalance

With SMOTE

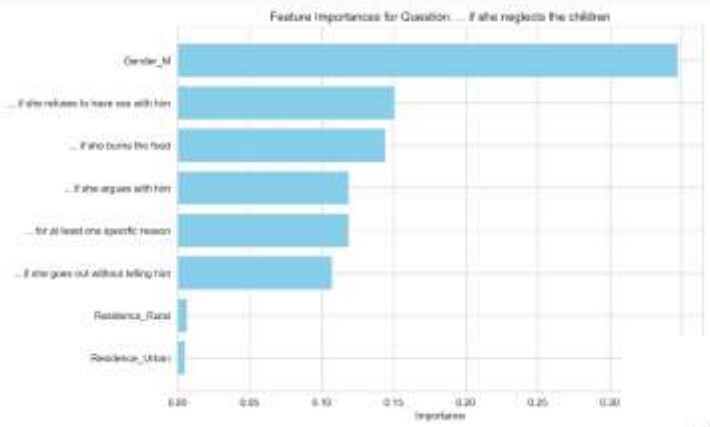
Accuracy remained 0.2047

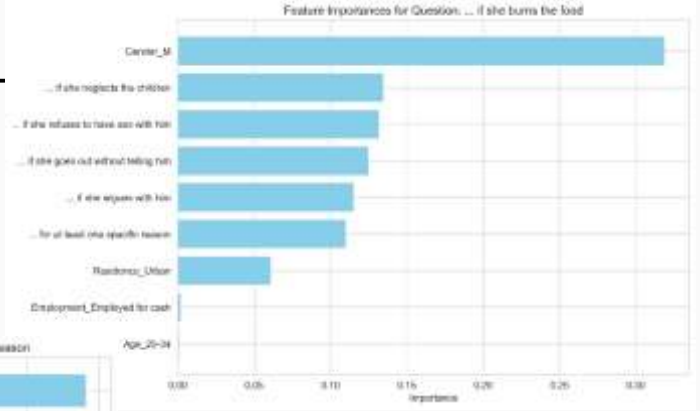
3 class target and SMOTENC

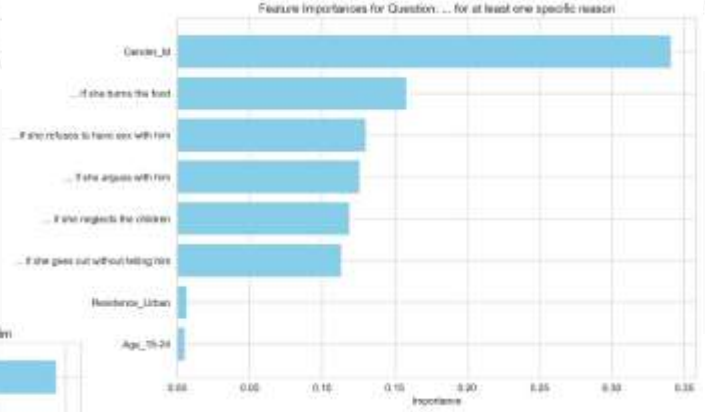
Survey responses were converted into 3 levels: 0(no justification), 1(minor justification) and 2(high)
Lower accuracy (0.148)

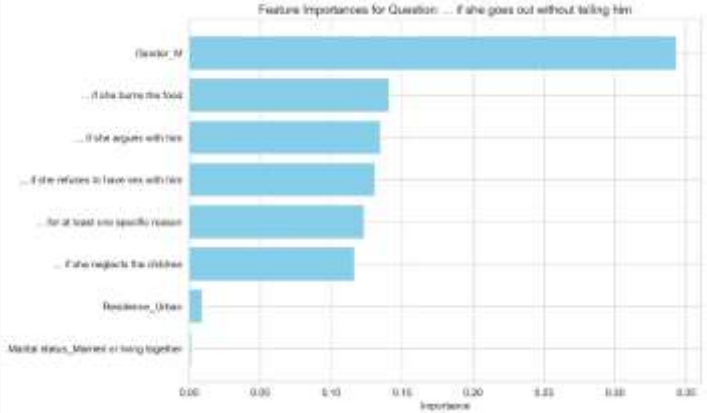
SMOTENC and all survey questions

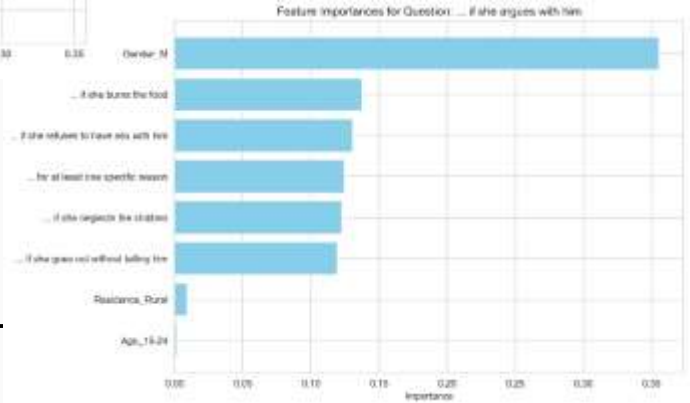
Models were trained using both demographic features and survey responses. SMOTENC was applied before one-hot encoding, which was applied only to demographic features


Feature Importances for Question: ... if she neglects the children
Gender_M
... if she refuses to have sex with him
... if she burns the food
... if she argues with him
... for at least one specific reason
... if she goes out without telling him
Residence_Rural
Residence_Urban
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Importance


Feature Importances for Question: ... if she burns the food
Gender_M
... if she neglects the children
... if she refuses to have sex with him
... if she goes out without telling him
... if she argues with him
... for at least one specific reason
Residence_Urban
Employment_Employed for cash
Age_25-34
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Importance


Feature Importances for Question: ... for at least one specific reason
Gender_M
... if she burns the food
... if she refuses to have sex with him
... if she argues with him
... if she neglects the children
... if she goes out without telling him
Residence_Urban
Age_15-24
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Importance


Feature Importances for Question: ... if she goes out without telling him
Gender_M
... if she burns the food
... if she argues with him
... if she refuses to have sex with him
... for at least one specific reason
... if she neglects the children
Residence_Urban
Marital status_Married or living together
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Importance


Feature Importances for Question: ... if she argues with him
Gender_M
... if she burns the food
... if she refuses to have sex with him
... for at least one specific reason
... if she neglects the children
... if she goes out without telling him
Residence_Rural
Age_15-24
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Importance

Summary

PCA
Feature loadings

Rural residence	60%
Age 15-24	20%
Employed	30%
Married	30%

Logistic regression
Feature coefficients

Rural residence	80%
Age 15-24	70%
Married	60%
Employed	40%

Decision trees
Feature importance

Male	30%
Other survey justification	10%
Rural residence	5%
Age 15-24	5%

Conclusion 07

The analysis demonstrates that demographic context plays a dominant role in attitudes toward IPV, with rural residence, youth, marital status, and employment status being consistent predictors across multiple approaches.

The findings strengthen the need for targeted, context-sensitive interventions that address both the structural demographic realities and the specific correlated beliefs driving the acceptance of violence.