



# Understanding attitudes, experience and risk factors for domestic violence

International Team of Scientists working at JAMSTEC, Japan



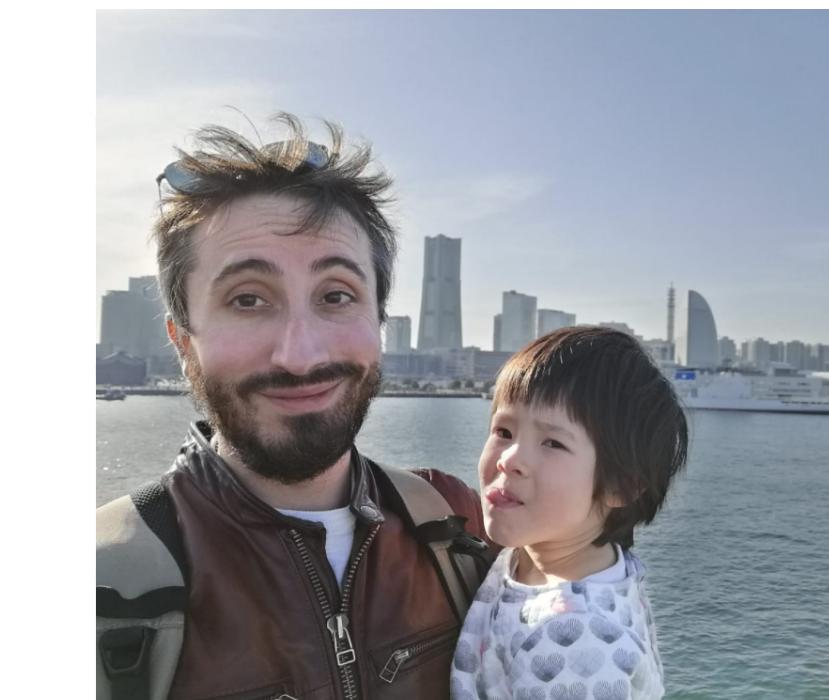
Prima Anugerahanti  
Indonesia



Gerlien Verhaegen  
Belgium



Heather Ritchie  
Scotland



Arthur Bauville  
France



Ettore Barbieri  
Italy

The Demographic and Health Surveys (DHS) Program has collected, analyzed, and disseminated accurate and representative data on population, health, HIV, and nutrition through more than **400 surveys in over 90 countries**.



*A mother and daughters in Jimma, Ethiopia work with coffee beans after their house has received Indoor Residual Spraying (IRS) to reduce malaria transmission. Photo Credit: AIRS Ethiopia PMI*

## Using data provided by the DHS Program we wanted to:

- ▶ Investigate the **acceptance vs the experience** of domestic violence
- ▶ Determine the **controlling factors** of domestic violence
- ▶ Develop a tool that would **quickly and efficiently identify women who are at risk** of domestic violence

The Demographic and Health Surveys (DHS) Program has collected, analyzed, and disseminated accurate and representative data on population, health, HIV, and nutrition through more than 400 surveys in over 90 countries.



A mother and daughters in Jimma, Ethiopia work with coffee beans after their house has received Indoor Residual Spraying (IRS) to reduce malaria transmission. Photo Credit: ALRS Ethiopia PMI

## DHS Program data for acceptance and experience of domestic violence

► Acceptance: % of yes to at least one answer of:

It is the respondent's opinion that a husband is justified in hitting or beating his wife when:

**BASE:** For V744A to V744E is all women.

- V744A She goes out without telling him
- V744B She neglects the children
- V744C She argues with him
- V744D She refuses to have sex with him

► Experience: % of yes to at least one answer of:

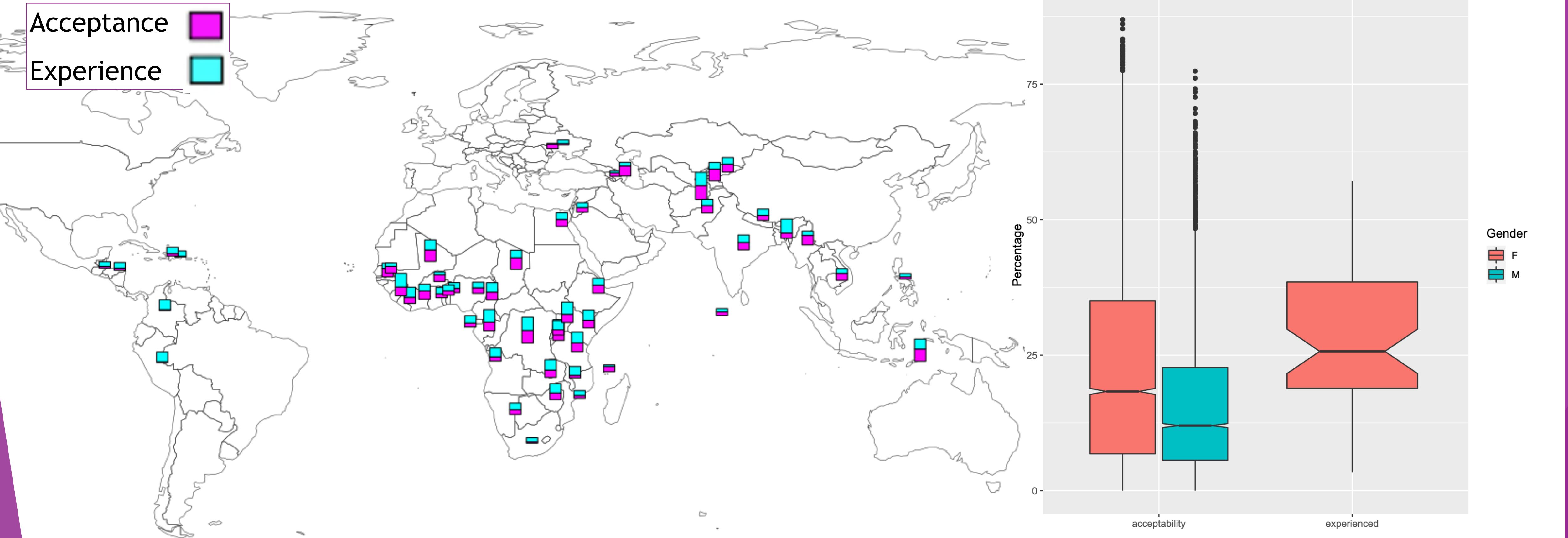
### Physical violence

- BASE:** Women currently in union ( $V502 = 1$ ).  
Spouse ever pushed, shook or threw something.  
Spouse ever slapped.  
Spouse ever punched with fist or something harmful.  
Spouse ever kicked or dragged.  
Spouse ever tried to strangle or burn.  
Spouse ever threatened with knife/gun or other weapon.  
Spouse ever attacked with knife/gun or other weapon.  
Spouse ever physically forced sex when not wanted.  
Spouse ever forced other sexual acts when not wanted.  
Spouse ever twisted her arm or pulled her hair.

- D105A
- D105B
- D105C
- D105D
- D105E
- D105F
- D105G
- D105H
- D105I
- D105J
- D105K
- D105L
- D105M
- D105N
- D106
- D107
- D108

- Physical violence
- Physical violence
- Physical violence
- Physical violence
- Experienced any less severe violence.
- Experienced any severe violence.
- Experienced any sexual violence.

# Acceptance vs experience of domestic violence

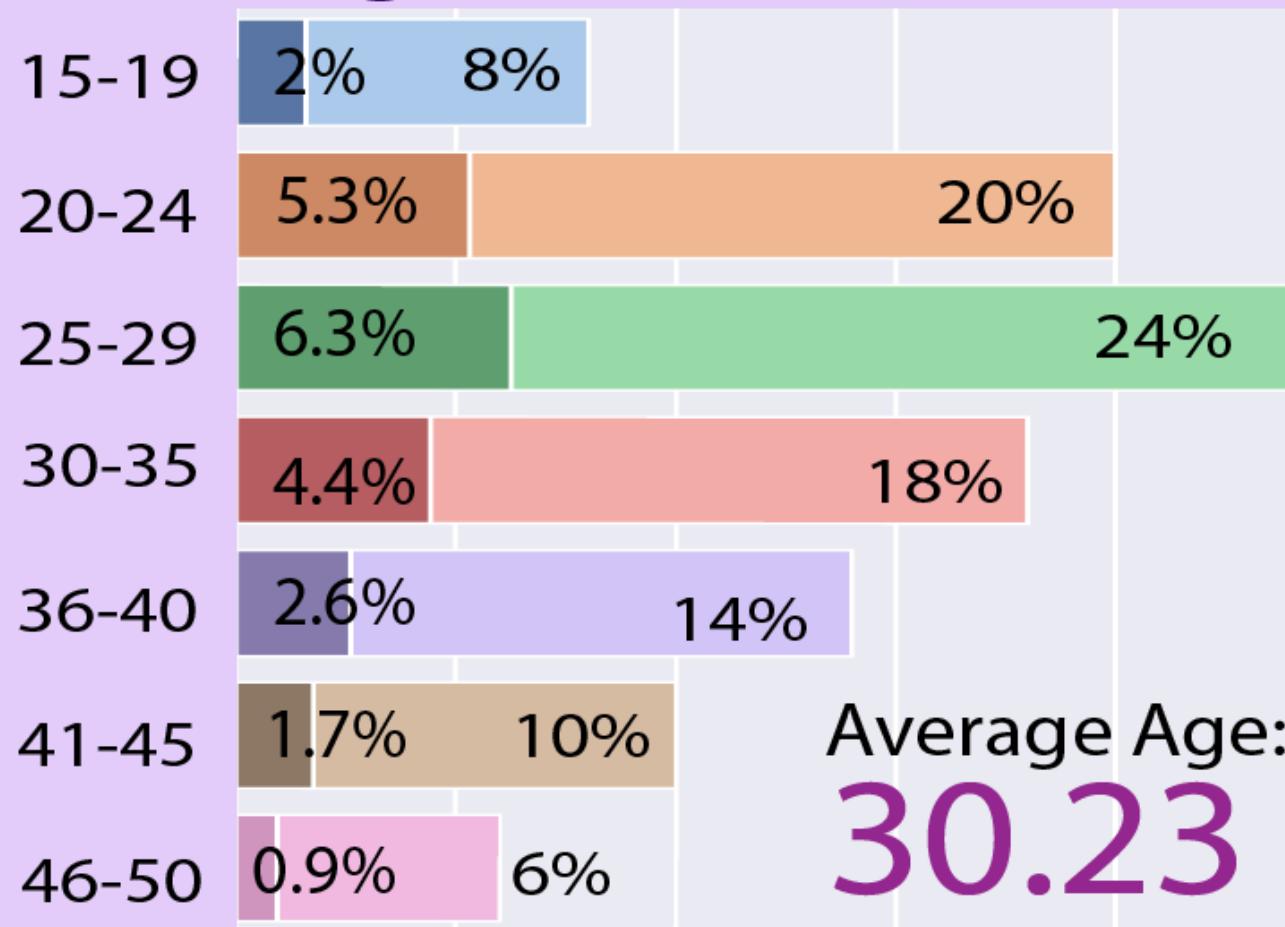


- ▶ Domestic violence is **widespread** amongst the countries surveyed by the DHS
- ▶ On average across all countries **26% of women** surveyed have experienced physical and/or sexual domestic violence
- ▶ Overall **more women than men** agreed that domestic violence was acceptable in certain circumstance

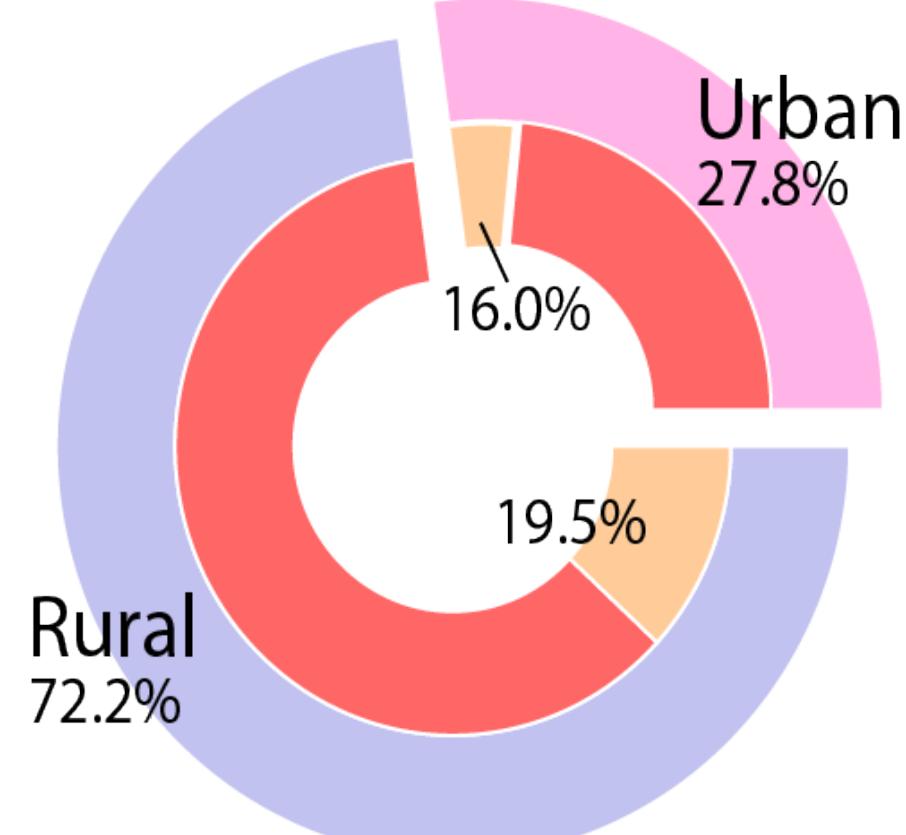
# Demographic of Respondents

5120 female and married respondents in Congo between 2013-2014

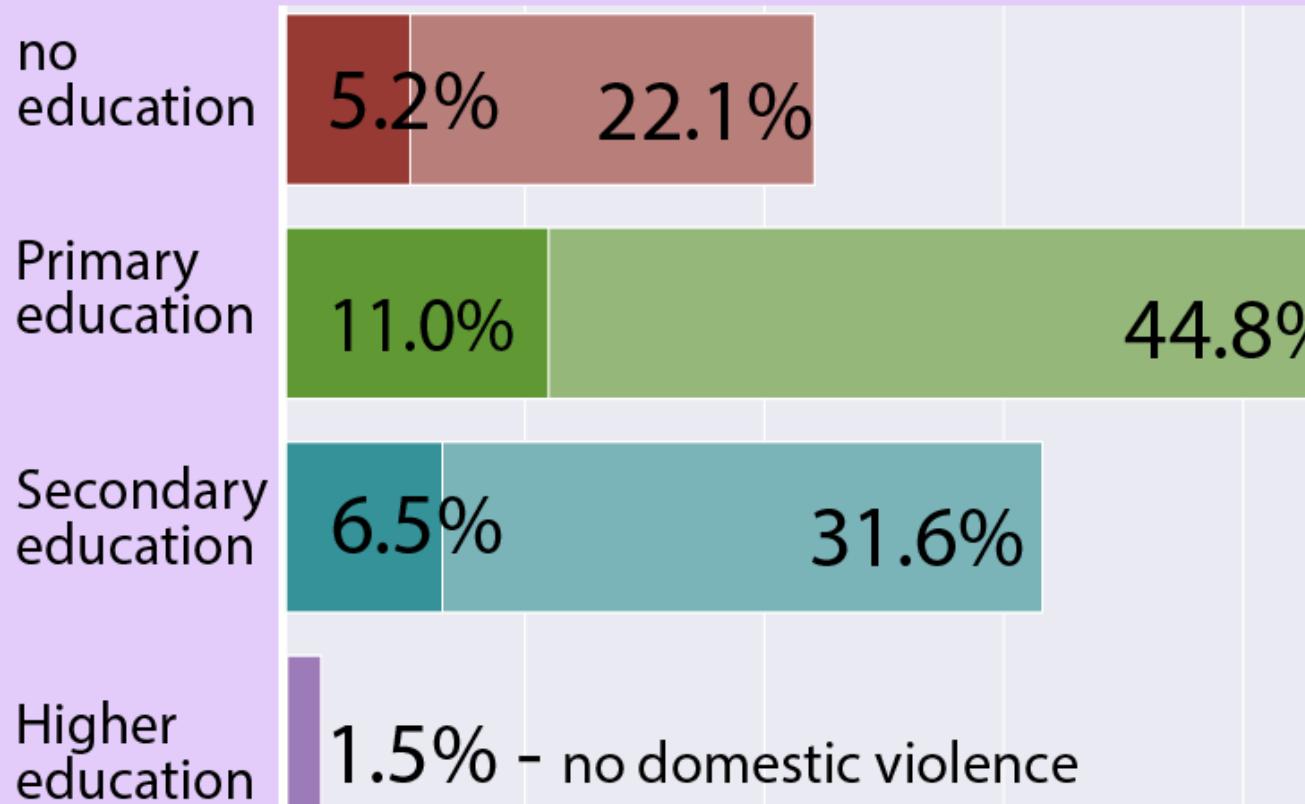
## Age



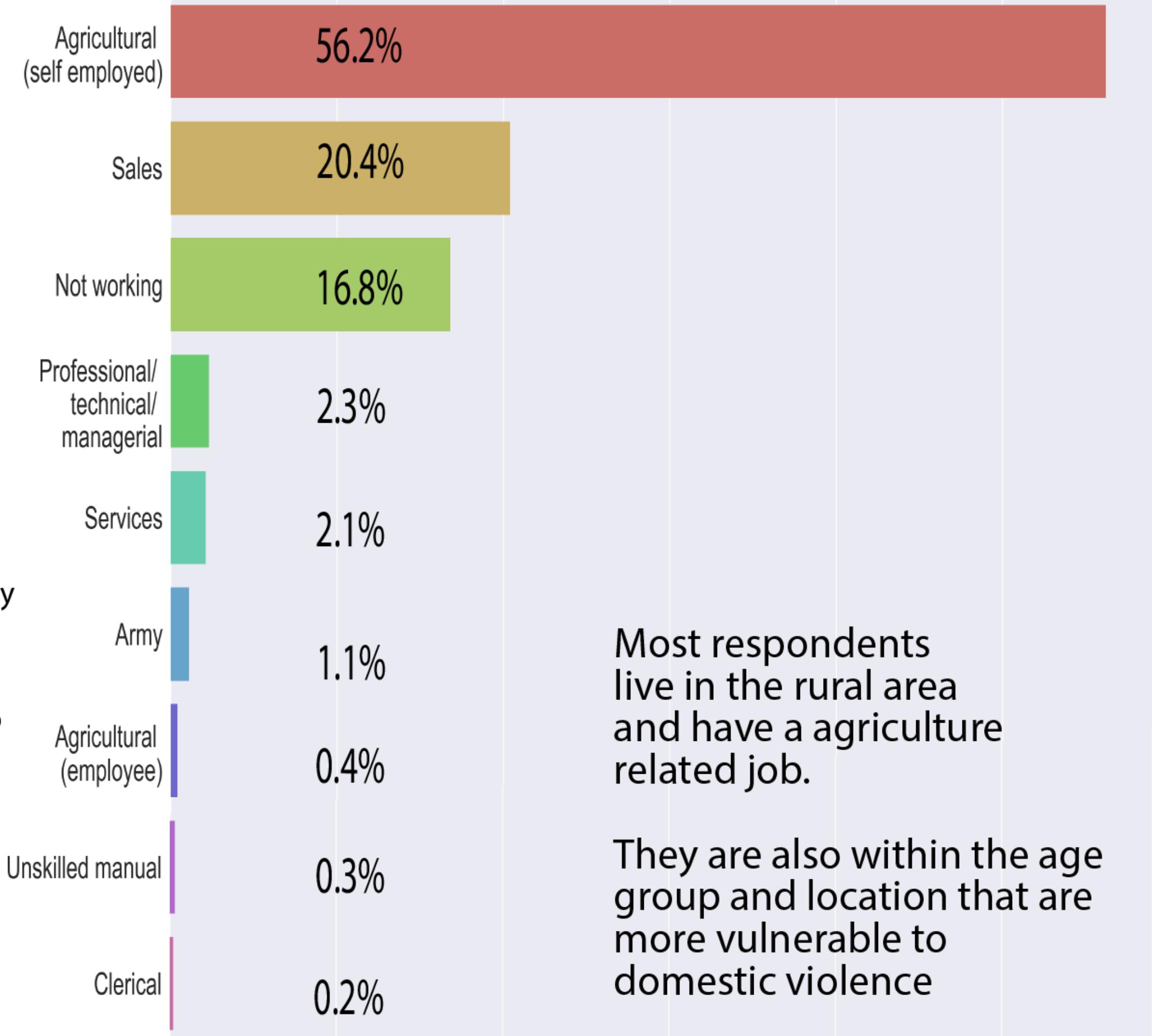
## Location



## Education



## Occupation

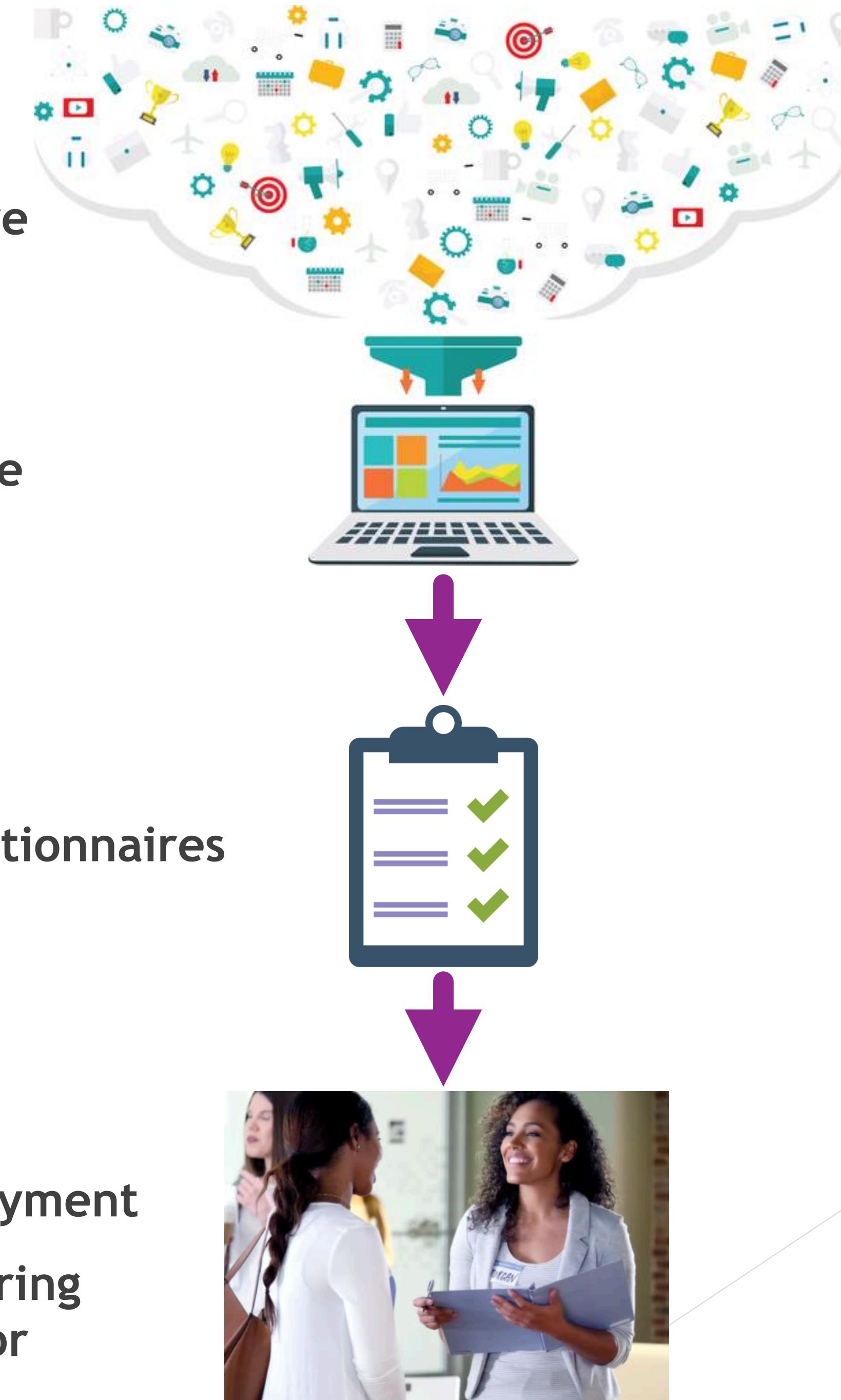


# Can we use these data to predict which women are at most risk of future violence?

- ▶ Domestic violence is a severe problem
  - ▶ Skilled professionals (social workers, academics, politicians, doctors) are working everyday on prevention and helping victims
  - ▶ Machine learning is **NOT** going to magically solve this problem...
  - ▶ ... but maybe we can provide an additional tool to these professionals
- ▶ Specifications:
  - ▶ implementable without computer (paper, dialogue) by a non-expert (to be deployed in developing countries)
  - ▶ Test under 2 minutes → the test can be taken during routine visit to the doctor or administration, even in the presence of the husband
  - ▶ Predictive (detection of women at risk of violence before it occurs)
  - ▶ General (ideally, applicable in many countries)

# Proposed solution

- ▶ DHS survey questionnaire
  - ▶ Binary classification tree
  - ▶ short personalized questionnaires
  - ▶ Paper or dialogue deployment
  - ▶ Quick detection test during routine administrative or medical visits



# Data

- ▶ DHS survey women's questionnaire (>70 pages, >500 questions)
- ▶ For a given country: ~10,000 individuals, ~4000 features
- ▶ Numerical, categorical, missing values
- ▶ The dataset required a certain amount of cleaning (missing values encoded in different format, categorical encoding etc...). Since time was limited we manually selected 38 questions (75 features) for data exploration including:
  - ▶ **background:** wealth index, age, number of children, education level, religion, etc...
  - ▶ **Husband background:** age, education level, use of alcohol, control issues
  - ▶ **Violence:** emotional, physical and sexual

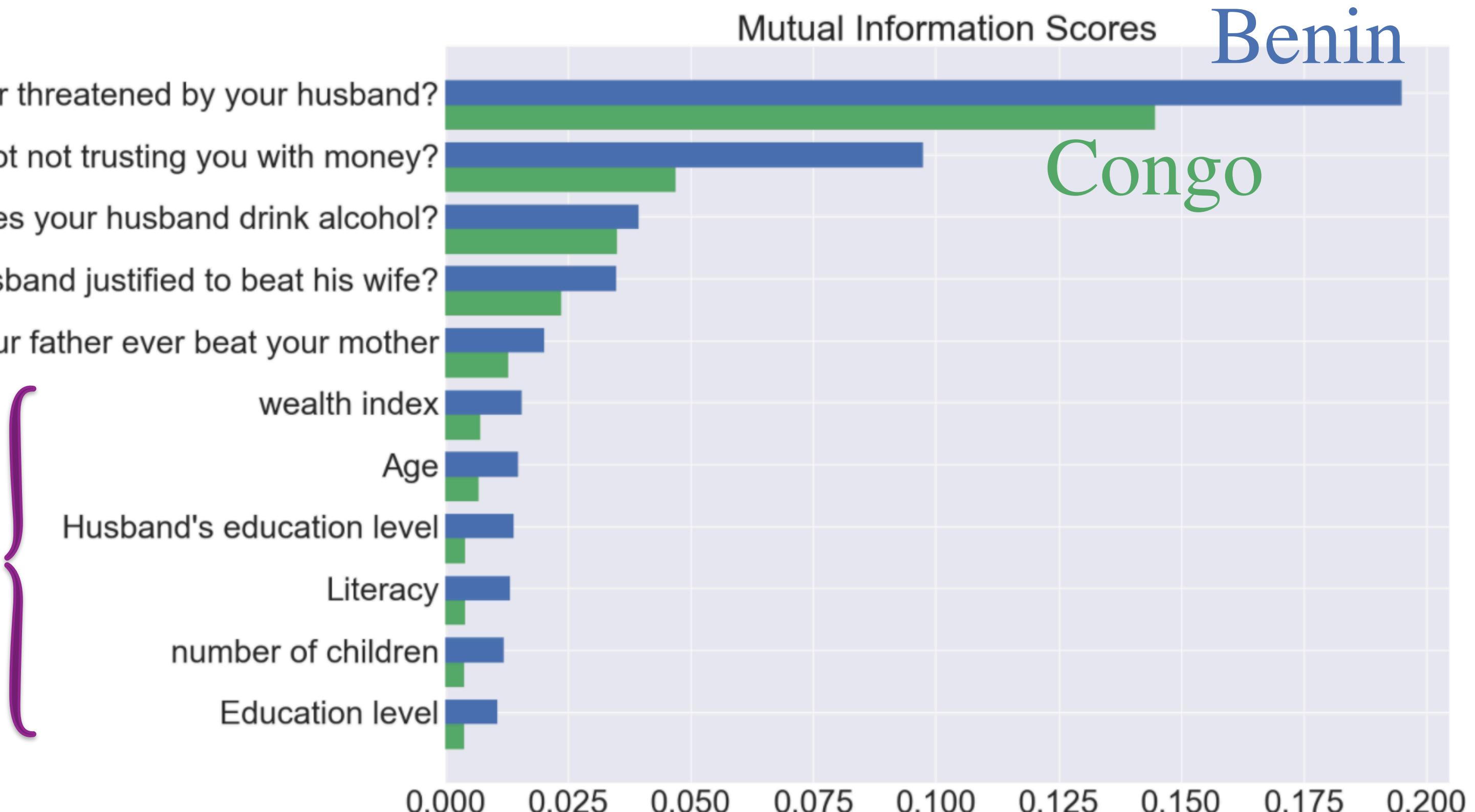
# Feature selection

From which other question can we predict answers to the question:

► “Have you experienced physical or domestic violence in the last 12 months?”

- Emotional violence:** Have you ever been humiliated, insulted or threatened by your husband?
- Control issues:** Is your husband jealous, or not trusting you with money?
- Alcohol consumption:** How often does your husband drink alcohol?
- Acceptance of violence:** Is a husband justified to beat his wife?
- Family history:** Did your father ever beat your mother

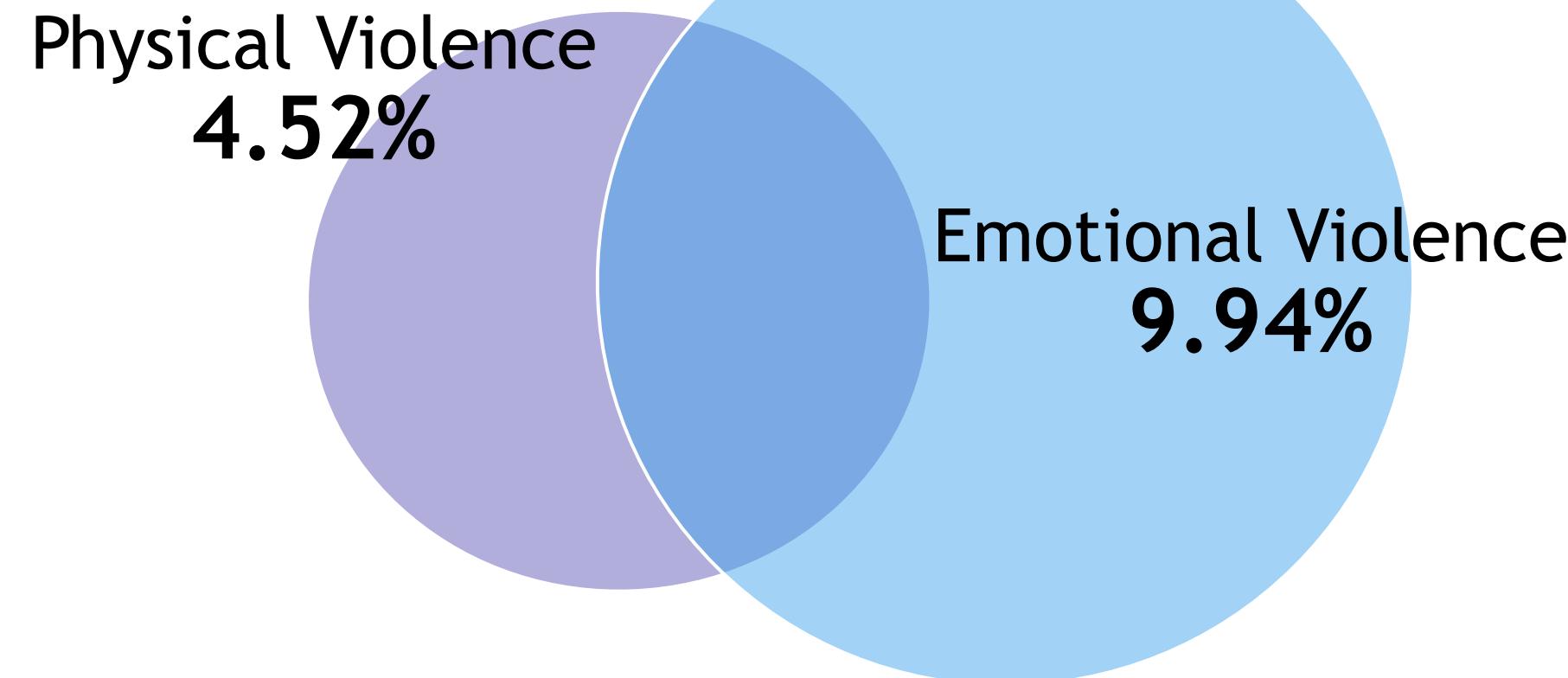
Demographics:



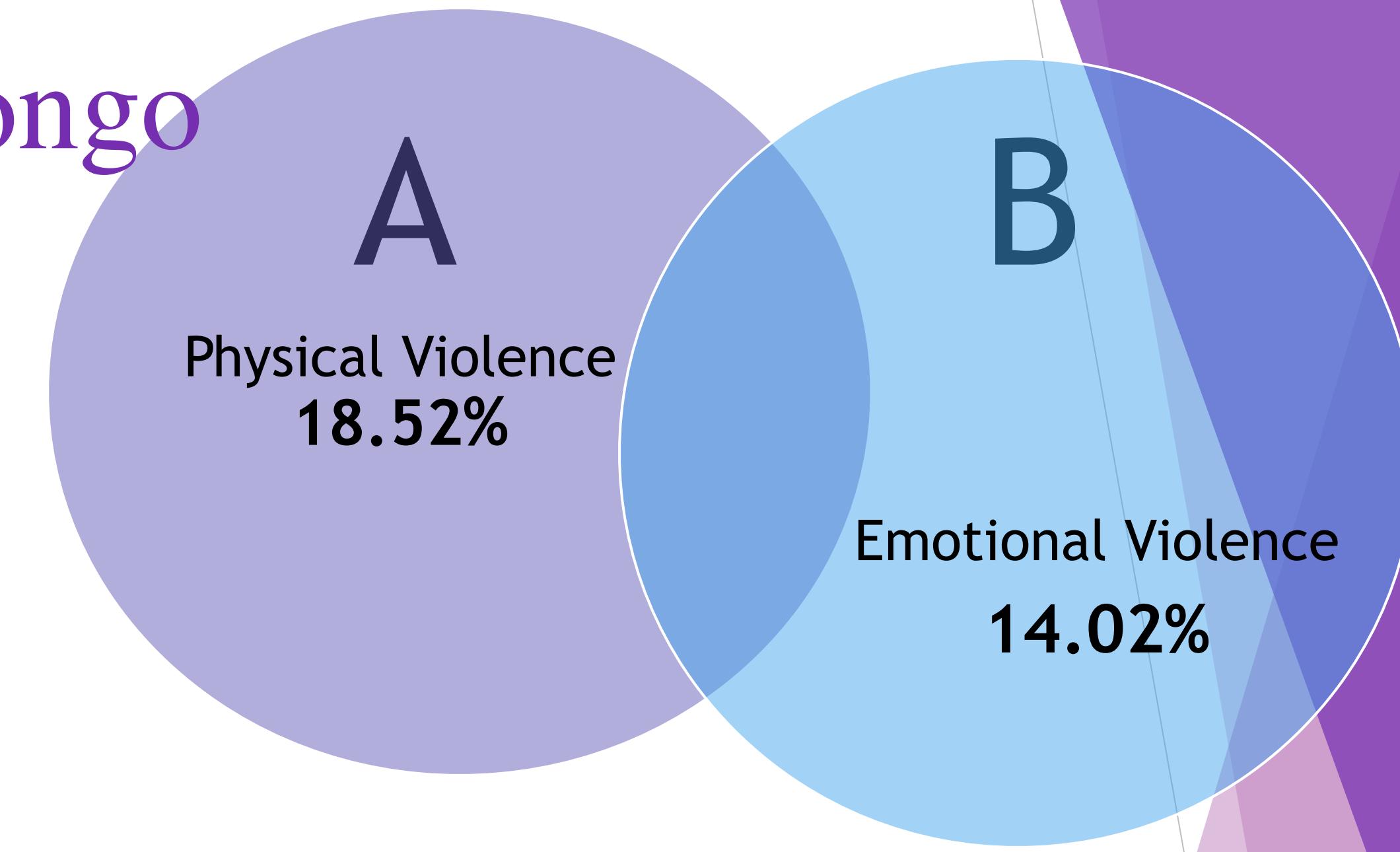
The relative importance of features is the same in Benin and Congo which suggests that the model is general (although more countries need to be tested to confirm)

# Is emotional violence predictive of physical/sexual violence

Benin



Congo



$$P(B | A)$$

$$\begin{cases} \text{(Benin)} & 63\% \\ \text{(Congo)} & 51\% \end{cases}$$

} of women experiencing **physical violence** also experience **emotional violence**

$$P(A | B)$$

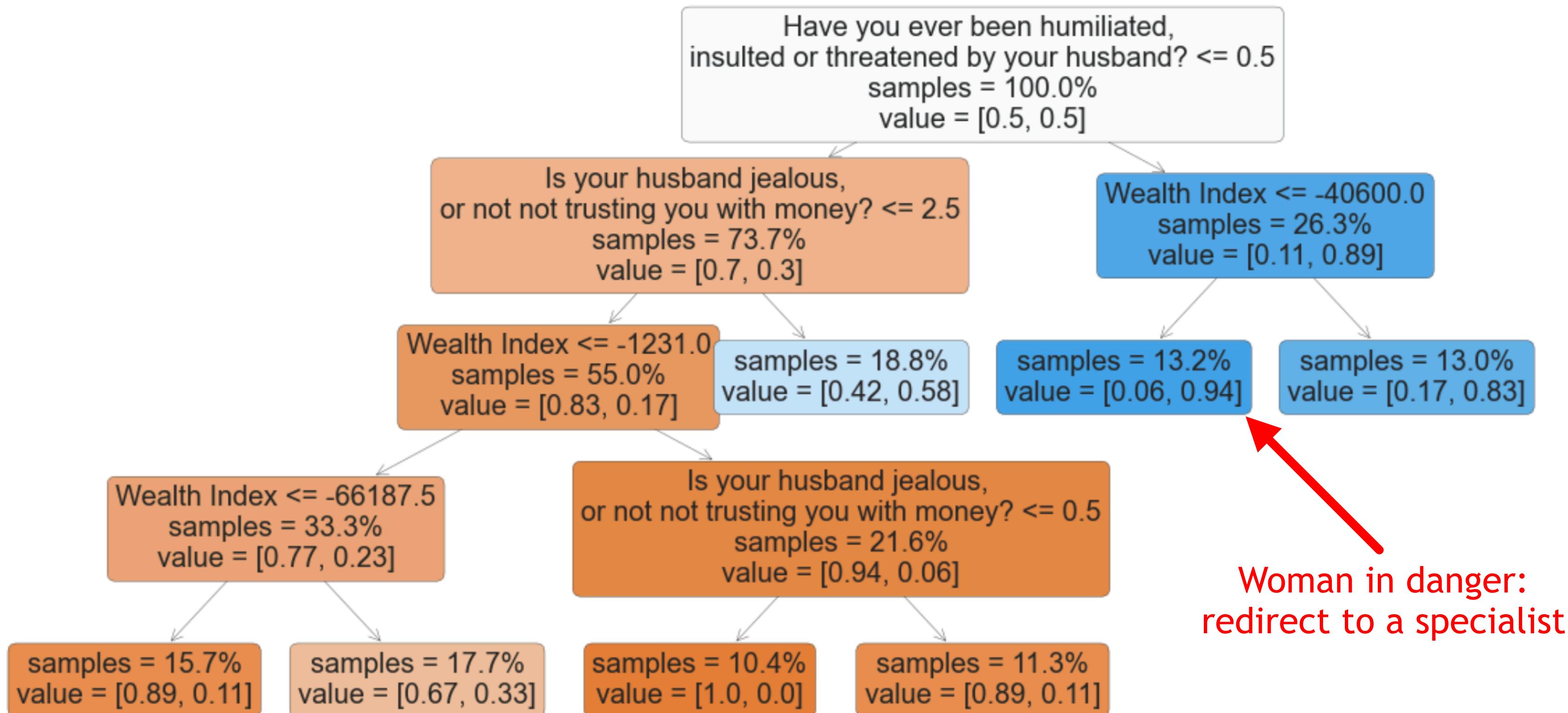
$$\begin{cases} \text{(Benin)} & 29\% \\ \text{(Congo)} & 67\% \end{cases}$$

} of women experiencing **emotional violence** also experience **physical violence**

In both countries emotional and physical violence are correlated but emotional violence seems to be a precursor of physical/sexual violence only in Benin

# Questionnaire derived from classification tree

- ▶ Binary classification tree: doesn't require a computer (paper, dialogue)
- ▶ True: this woman is experiencing violence



Woman in danger:  
redirect to a specialist

# Model performance

- ▶ True: this woman is experiencing violence
- ▶ Since the dataset was initially imbalanced we subsampled the majority class to reach 50% False, 50% True
- ▶ Hyperparameters were tuned to optimize **recall** because:
  - ▶ True positives are women in danger (experiencing physical or sexual violence)
  - ▶ False negatives are women in danger but not detected (**should be minimized**)
  - ▶ False positives are women who are not experiencing physical violence but may in the future

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

		Predicted label		Model's performance				
		False	True	Metrics				
Actual	False	38%	10%	precision	recall	f1-score	support	
	True	13%	38%					

	False	True	precision	recall	f1-score	support
False	0.74	0.78	0.74	0.78	0.76	37
True	0.78	0.74	0.76	0.74	0.76	39

# Limitations of the present results

- ▶ Decision trees are not robust such that a slight change in the dataset leads to a different tree. Ensemble method (e.g. random forest) or not desirable here because the model would become a black box and necessitate a computer.
- ▶ We used only a small subset of features
- ▶ Emotional violence, husband control issues or drinking habits may be delicate subjects.
  - ▶ Those are the best questions to ask your family or friends
  - ▶ However, when asked by a unknown person these subjects may seem intrusive during such a short questionnaire
- ▶ Therefore, we could select features that include less personal questions

# Conclusion

- ▶ Although domestic violence is often not considered acceptable it is still widespread
- ▶ We proposed to use machine learning to design personalized questionnaire to assist in the detection of women at risk of domestic violence
- ▶ Our results suggests that emotional violence, control issues and alcohol consumption are significant factors in the occurrence of domestic factors.
- ▶ The questionnaire contains rather intimate questions which limits its usability. The next challenge would be to identify new features that allow for a less intrusive questionnaire while maintaining good model performance.







# Additional information

## Data

## Problematics

## Skills tested

## Winner profile

## Kaggle

Given data is clean  
only given data allowed

Given problem, goal, metric

Technical: Feature engineering,  
Modeling

Professional ML engineer

## Datathon

Given data may not be clean  
other data allowed

Given theme, open problematics

Problem formulation,  
Communication: Visualization, presentation,  
Technical: statistics, modeling, deployment

ML engineer, data analyst, researcher (any field), CS student

# Plan

- ▶ Feature selection
- ▶ Is the model general?
- ▶ Is the model predictive?
- ▶ Example questionnaire
- ▶ Model performance

# Can we use these data to predict which women are at most risk of future violence?



# Model specifications



- ▶ The dataset treats of developing countries. In order to deploy the model in these countries we imposed ourselves the following specifications:
  - ▶ short personalized questionnaire
  - ▶ implementable without computer by a non-expert
  - ▶ Test under 2 minutes (less than 5 questions) -> the test can be taken during routine visit to the doctor or administration, even in the presence of the husband
  - ▶ Predictive
  - ▶ General
- ▶ To comply with these specifications we chose a **binary classification tree**
- ▶ Which features to use and how predictive are they?
- ▶ Is the model general (applicable worldwide) or specific (to a given region)
- ▶ Is the model predictive? i.e. can we detect only identify women already experiencing physical violence or can we detect women at risk before violence occurs?

# Model specifications

- ▶ The DHS program questionnaire is very long (>50 pages). It cannot be performed casually, but it may already contain the right questions for a quick diagnostic
- ▶ We use a decision tree classification model:
  - ▶ short personalized questionnaire
  - ▶ implementable without computer by a non-expert (i.e. deployable in developing countries)
  - ▶ Test under 2 minutes (less than 5 questions) -> the test can be taken during routine visit to the doctor or administration, even in the presence of the husband
- ▶ Ideally, the model should be universal and predictive
- ▶ To assess the usability of the model we need to determine three things:
- ▶ Which features to use and how predictive are they?
- ▶ Is the model general (applicable worldwide) or specific (to a given region)
- ▶ Is the model predictive? i.e. can we detect only identify women already experiencing physical violence or can we detect women at risk before violence occurs?



# Data

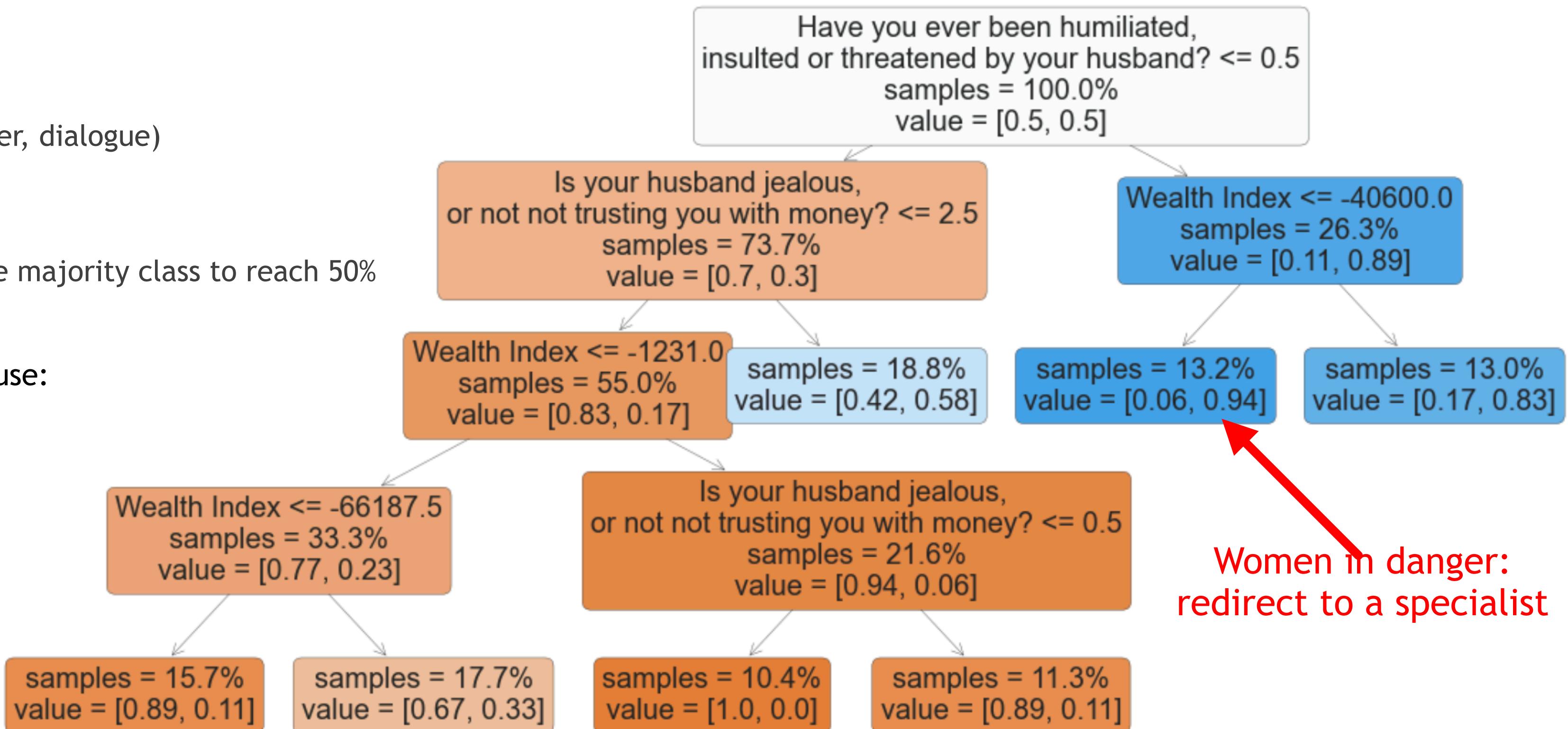
- ▶ DHS survey women's questionnaire (>70 pages, >500 questions)
  - ▶ **Background characteristics**
  - ▶ **Husband's background:** Currently married women are asked about the age, education, and occupation of their husbands.
  - ▶ **Status of women:** The questionnaire asks about various aspects of women's empowerment, including decision making, autonomy, ownership of houses and land, barriers to medical care, and attitudes towards domestic violence.
  - ▶ Other topics: Questions examine behavior related to environmental health, the use of tobacco, and health insurance.
- ▶ ~10,000 individuals, 4275 features
- ▶ Numerical, categorical, missing values
- ▶ Here, we present results with questionnaire in Congo and Benin

# How well can we identify women at risk of physical violence?

- ▶ Binary classification tree: doesn't require a computer (paper, dialogue)
- ▶ True: this woman is experiencing violence
- ▶ Since the dataset is initially imbalanced we subsampled the majority class to reach 50% False, 50% True
- ▶ Hyperparameters were tuned to optimize **recall** because:

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

- ▶ True positives are women in danger (experiencing physical or sexual violence)
- ▶ False negatives are women in danger but not detected (**should be minimized**)
- ▶ False positives are women who are not experiencing physical violence but may in the future



```

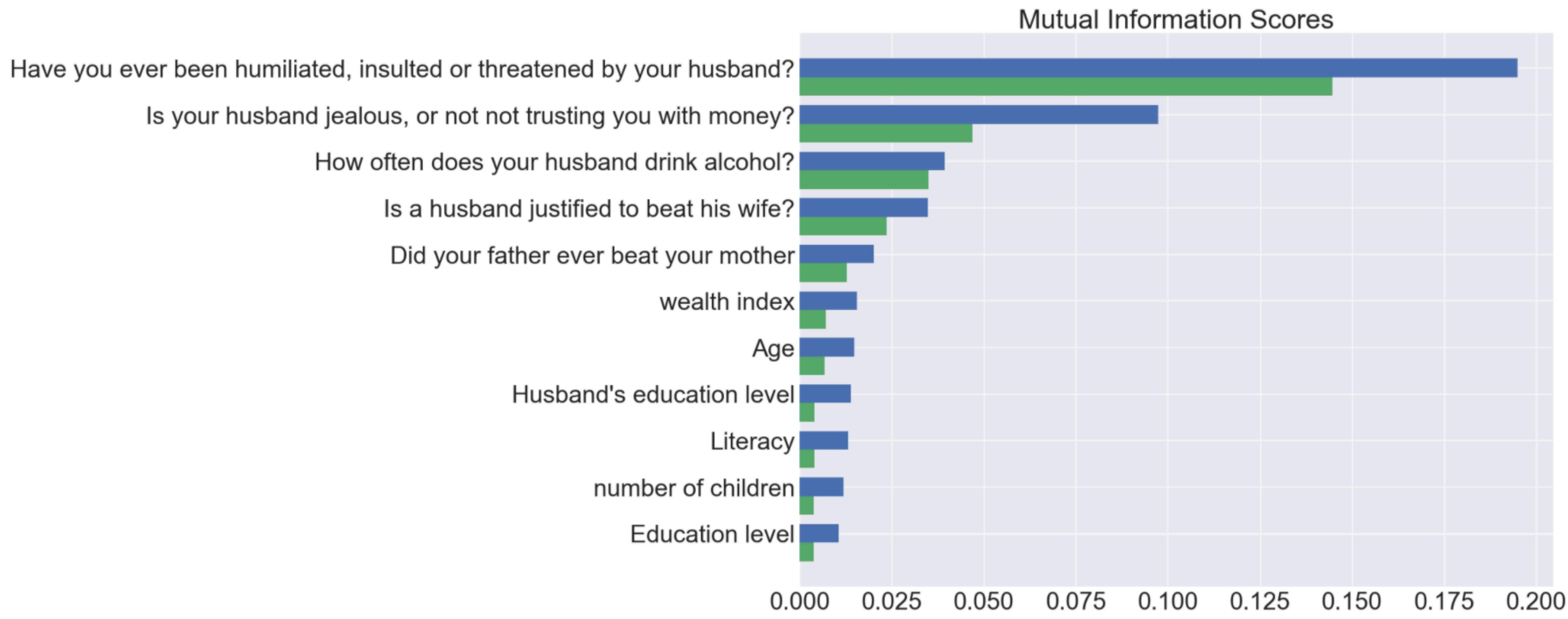
for itarget, target_feature in enumerate([physical_violence]):
    y = X[target_feature + '_sum']>0

    plt.subplot(1,2,itarget+1)
    this_X = X.drop(columns=[target_feature + "_sum"] )
    discrete_features = this_X.dtypes == 'int8'
    mi_scores = pd.Series(np.zeros(this_X.columns.shape), index= this_X.columns)
    n = 20
    for i in range(n):
        X_sub, y_sub = subsample(this_X,y)

        mi_scores += 1.0/n*make_mi_scores(X_sub, y_sub, discrete_features)

mi_scores.rename(dict(zip(X.columns,explicit_feature_names)), inplace=True)
plot_mi_scores(mi_scores)

```



# Model code

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import GradientBoostingClassifier, RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn import tree
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import f1_score, roc_auc_score
y = X[physical_violence + "_sum"]>0
this_X = X.drop(columns=[physical_violence + "_sum"])
this_X.columns
discrete_features = this_X.dtypes == 'int8'
X_model = pd.get_dummies(this_X, columns=this_X.columns[discrete_features])
imbalance_fac = 1
X_model, y = subsample(X_model,y, imbalance_fac=imbalance_fac)
X_train, X_test, y_train, y_test = train_test_split(X_model.values,y.values, test_size=0.2)

param_grid = dict(max_depth=[9],
                  min_samples_leaf=[0.05, 0.07, 0.1, 0.15, 0.2, 0.25, .3, 0.5],
                  )

clf = DecisionTreeClassifier(min_impurity_decrease=0.0001,
                             class_weight={0:1, 1:imbalance_fac},
                             )

clf_cv = GridSearchCV(clf, param_grid, cv=3, scoring='recall')

clf_cv.fit(X_train, y_train)
display(clf_cv.best_params_)
y_pred = clf_cv.best_estimator_.predict(X_test)

print(confusion_matrix(y_test, y_pred)/y_test.shape[0])
classif_dict = classification_report(y_test,y_pred, output_dict=True)
print("f1_score:", classif_dict['True']['f1-score'])
print(classification_report(y_test,y_pred))
```

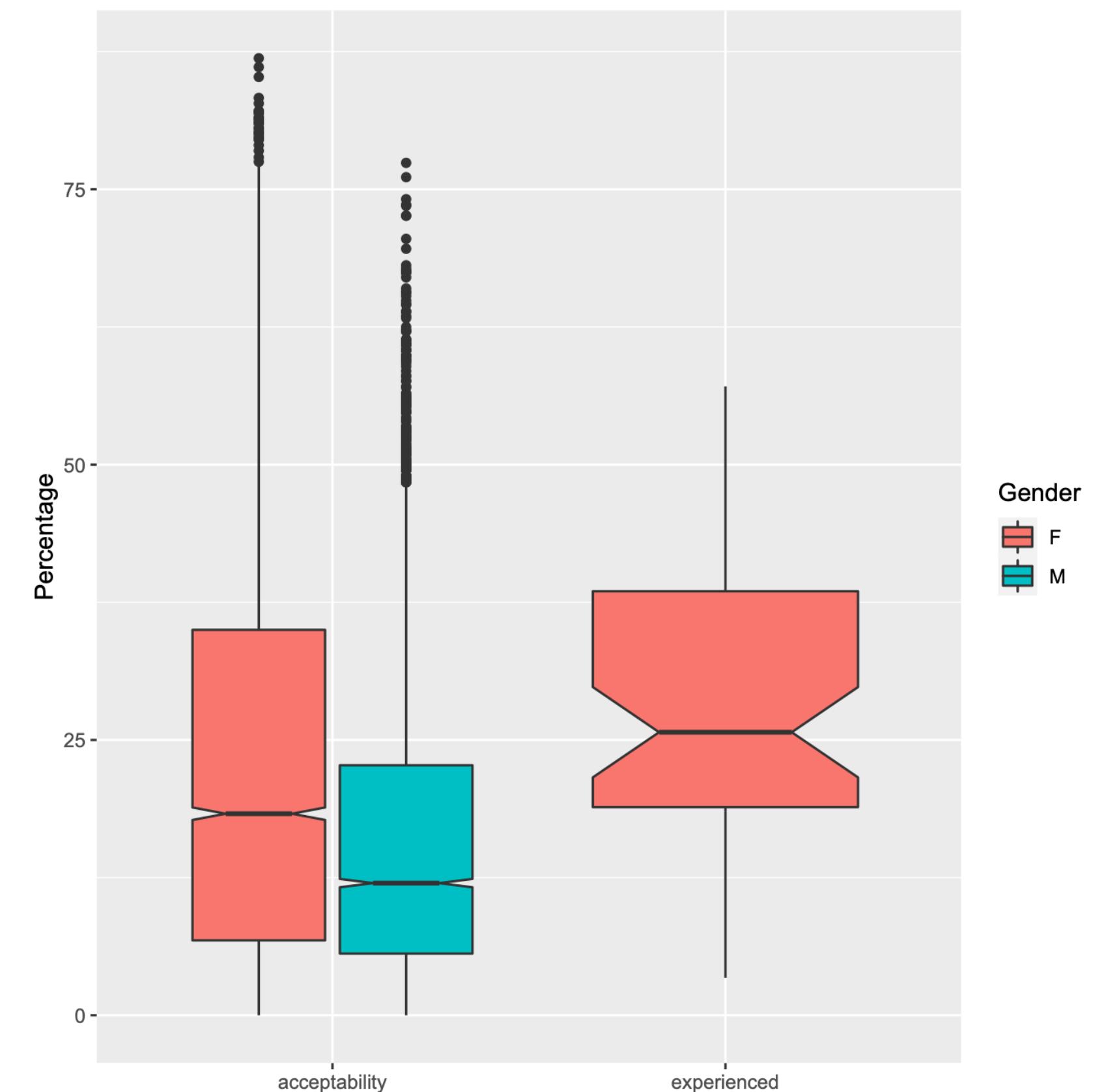
# Can we use these data to predict which women are at most risk of future violence?



# Code

```
dat <- read.csv("acceptability_vs_experienced_vs_gender.csv")
library(ggplot2)

ggplot(dat, aes(x = Category, y = Value))+
  geom_boxplot(aes(fill = Gender), notch=TRUE, las=2)+
  labs(y="Percentage", x="")
```



# Code

```
library(rworldmap)  
  
mapBars( dat, nameX="lon", nameY="lat",  
         nameZs=c('acceptability','experienced'), barWidth = 3,  
         borderCol= "grey17", zColours = c("magenta","cyan"))
```

