Predicting Lifetime Physical and/or Sexual Intimate Partner Violence (IPV) against Women

Junior Ovince

Capstone Project – DATS6501

Fall 2019

Contents

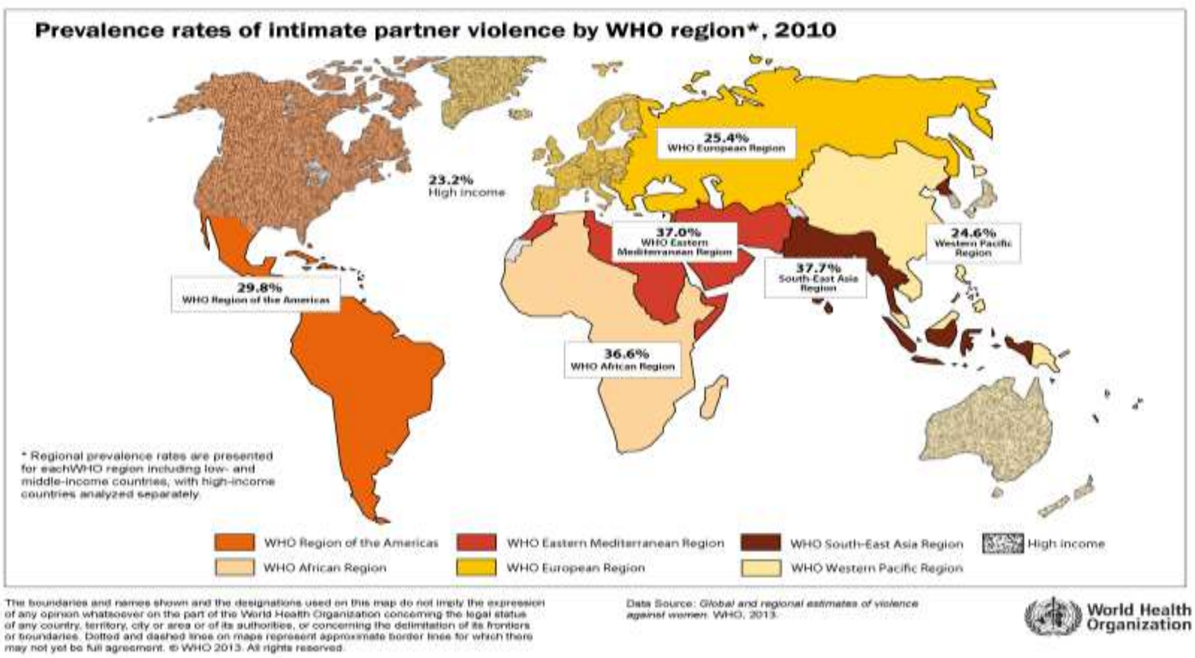
List of tables and figures

Problem definition

VAWG is a major public health problem and a crucial women's human right violation. In a recent report published in 2013, the World Health Organization (WHO) presented for the first time the global and regional estimates of intimate partner and non-partner violence and their health consequences. Here is a summary of some of the findings for the estimates of violence:

- A large proportion of the women's world, 35 per cent, have experienced either physical and/or sexual IPV or NPSV (see figure 1).
- Most of this violence are perpetrated by an intimate partner. Worldwide, one out of three ever-partnered women, 30 per cent, have suffered physical and sexual IPV. In some regions, this rate can as high as 38 per cent.
- Globally, in 38 per cent of a woman's murder case, the perpetrator is an intimate partner.
- A proportion of 7 per cent women worldwide have been sexually assaulted by someone other than a partner.

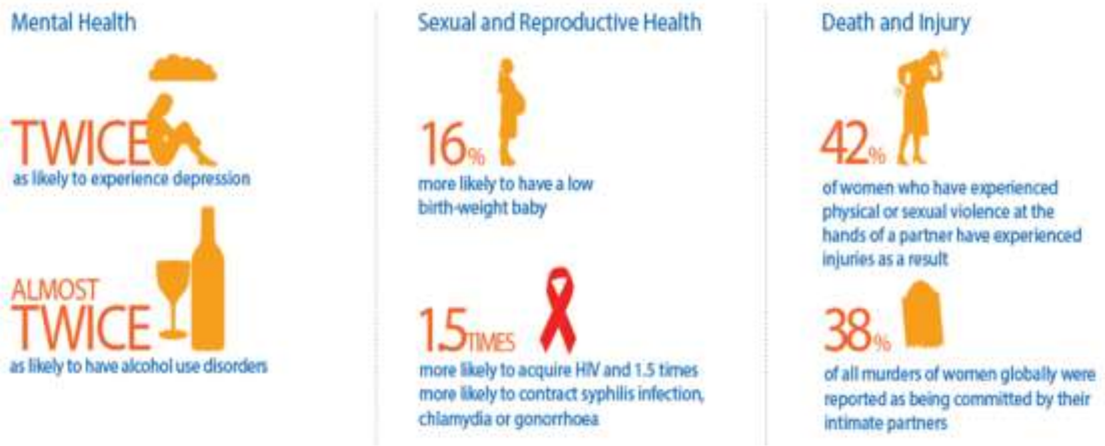
Figure 1. Worldwide Prevalence Rates of IPV



The health consequences of physical and/or sexual IPV are alarming (see figure 2):

- Low birth weight (16 per cent)
- Induced abortion (twice more)
- Depression (twice more)
- Incident of HIV infection (1.5 times more)
- Death and injuries
- Substance use disorders

Figure 2. Association between Intimate Partner Violence and Selected Health Outcomes[1]



A range of actions and efforts has been carried out both to prevent IPV from happening in the first place and to provide necessary services for survivors. Those actions have been informed or assessed by many studies and evaluations. Some of the research focus on determining the potential risk and preventive factors, while others aim to detect the impact/effect of a program or intervention. There is an opportunity to use machine learning algorithms to inform preventive and responsive actions to reduce physical and/or sexual IPV. In the first hand, such algorithms will inform on the most important features that contribute to IPV and existing or new preventive interventions can be accordingly designed. In the second hand, these machine learning techniques can be used to identify and better target potential survivors of violence in case of a responsive interventions.

Motivation, Objectives, and Research Questions

In the past two years, the UN Women, the Inter-American Development Bank (IADB), and the Global Women's Institute (GWI) have supported local statistical offices in different countries in the Caribbean region to conduct a large and representative national population survey on women's health and life experience using the WHO questionnaire and methodology. These studies aimed mainly to determine the national prevalence rate of IPV and NPV. Data were collected in five countries such as Jamaica, Trinidad and Tobago, Suriname, Guyana, and Grenada for women aged between 15 and 64 years old.

While the GWI and IADB will develop a report on the risk and protective factors for these countries using inferential analysis, this project aims to achieve two overarching objectives:

- Design and use a pattern recognition algorithm to classify survivors of violence

[1] This figure is originally produced by the WHO.

- Improve the overall recall and f1-scores of such algorithm to reduce false negative.

To achieve the aforementioned objectives, two research questions will be followed:
- What are the main factors from the ecological framework[2] that contribute the most to identify potential victims of IPV (features selection and extraction)?
- Which pattern recognition models will provide the best performance in classifying potential survivors of IPV?

Dataset Description

In this section, the target and the features of the dataset will be briefly described.

Target

The main target of the study is IPV (physical and/or sexual), and it is defined as self-reported experience of one or more acts of physical and/or sexual violence by a current or former partner since the age of 15 years (WHO definition).
- Physical violence is defined as: being slapped or having something thrown at you that could hurt you, being pushed or shoved, being hit with a fist or something else that could hurt, being kicked, dragged or beaten up, being choked or burnt on purpose, and/or being threatened with, or actually, having a gun, knife or other weapon used on you.
- Sexual violence is defined as: being physically forced to have sexual intercourse when you did not want to, having sexual intercourse because you were afraid of what your partner might do, and/or being forced to do something sexual that you found humiliating or degrading.

Table 1. Summary of the Data from the Caribbean National Surveys

	All	**Ever (Number)**	**Ever (Percentage)**	**IPV Prevalence**
Grenada	1076	987	91.7%	29.0
Guyana	1498	1391	92.9%	37.8
Jamaica	1070	975	91.1%	27.8
Trinidad and Tobago	1079	1019	94.4%	30.1
Suriname	1527	1423	93.2%	33.7
Total	**6250**	**5795**	**92.7%**	**32.3**

Features

All of the studies in violence against women and girls (VAWG) use the ecological model as a framework to determine women's and girls' risk factors of suffering IPV or other types of violence. The risk factors identified are grouped into four levels: individual, relationships,

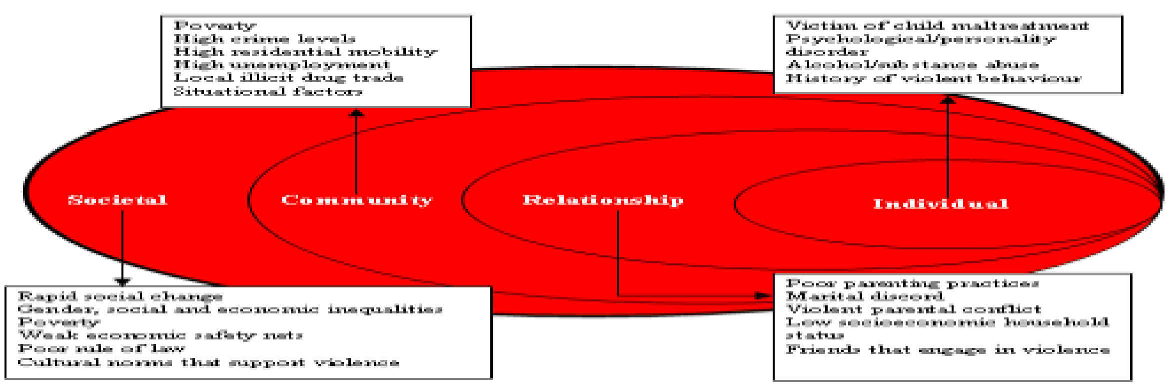
[2] http://www.endvawnow.org/en/articles/1509-the-ecological-framework.html

community, societal (see figure 3). The variables used as features in this study are mostly concentrated in the individual and relationships level (see table 2). Other variables such as economic and emotional violence are used to improve the performance of the models. In addition, some country level variables were defined to be included only in the multilevel regression models. All of the variables are categorical (binary or nominal/ordinal). The following table highlights the main components of the features based on the ecological model. A detailed table of each feature (48 in total) and the target is available in annex (table A1).

Table 2. Overview of the Variables according to the Ecological Framework

Category	**Examples**	**Levels**
Target	lifetime physical and/or sexual IPV (1 variable)	binary
Women's socio-demographic characteristics	age, religion, education, ethnicity, economic activities (7 variables)	individual
Women's social support	family support (1 variable)	relationship
Women's parenting and child abuse	childhood physical and emotional abuse, witnessing of violence (3 variables)	individual
Marital traditional norms	forced marriage, early marriage (2 variables)	societal
Partner's individual characteristics	age, education, economic activities (4 variables)	relationship
Partner's parenting and child abuse	childhood physical abuse, witnessing of violence (2 variables)	relationship
Partner's other behavior	extra-marital relationships, involvement in fight, use of alcohol (4 variables)	relationship
Couple relationship dynamics	couple communication, controlling behaviors (6 variables)	relationship
Women's attitude and gender norms	acceptance of traditional norms (11 variables)	community
Other violence	economic and emotional IPV (2 variables)	other
Country level variables	income, governance, rule of law, island, human development index (6 variables)	Country level

Figure 3. The Ecological Framework for VAWG[3]

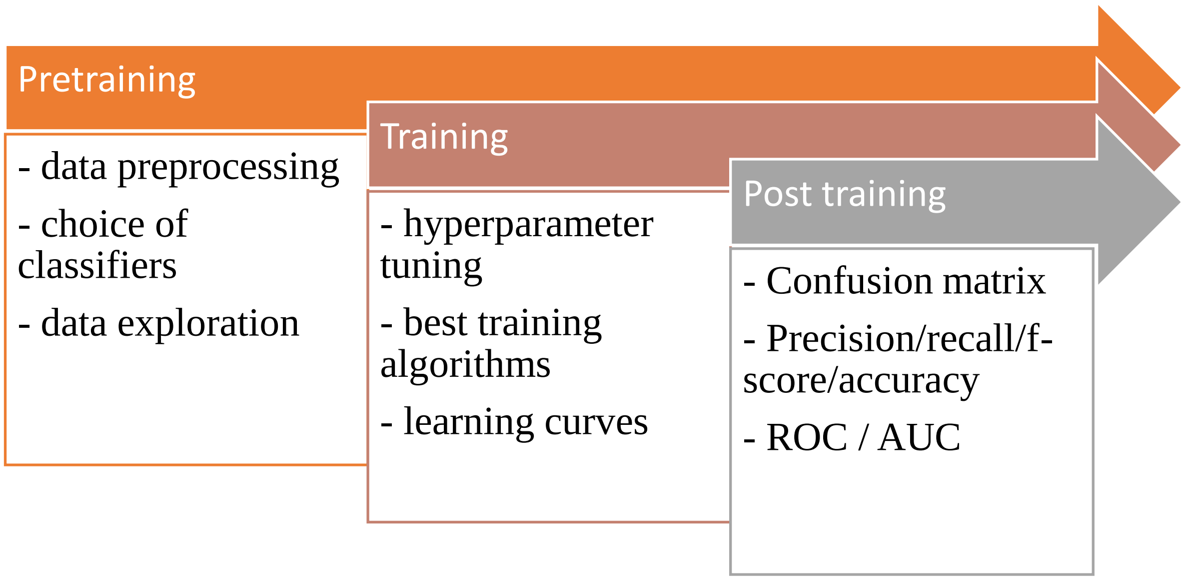


[3] https://www.who.int/violenceprevention/approach/ecology/en/

Methodology

The overall methodology can be divided into three main categories: pretraining, training, and post training. The pretraining phase serves as a basis for all the training phases, which differ based on the groups of classifiers in training. In the post training, the usual classification metrics to identify the best models that will be testing. This phase can be considered as the evaluation of the performance of the selected models before the final selection.

Figure 4. Overall Methodology Adopted



A. *Pretraining*

This phase includes three major components: the data preprocessing, the choice of the different supervised learning classifiers, and some exploratory analysis.

Data preprocessing

Many tasks were undertaken during the data preprocessing:

- Cleaning and recoding: All of the cleaning and recoding were done in SPSS as all the datasets were available in sps format. Most of the variables are recoded to ensure data comparison among the countries. The continuous variables such as age are recoded into categorical ones. In addition, other countries variables such as governance[4], Income, Human Development Index, and if the country is an island or not were included. All of the data for these variables are from the World Bank website.
- Checking for missing values: None of the variables have missing values based on a quick check on python/pandas.

[4] Voice and Accountability as well as rule of laws were considered. See this link for the methodology: https://info.worldbank.org/governance/wgi/#home

- One-hot encoding of nominal/ordinal variables: All of the variables are categorical variables. Most of them are binary ones and the few nominal or ordinal ones are recoded into binary variables.
- Splitting the data into train, validation, and test sets: The holdout cross-validation strategy is used to evaluate the best models selected. The data are divided into 70% for the train set and 15% each for the validation and test sets. Only the validation set is used during the model training.

Table 3. Summary of the Holdout Cross-validation Method Results

Set	N	%	IPV prevalence
Train	4054	69.957%	32.28
Validation	869	14.996%	32.22
Test	872	15.047%	32.22
Total	**5795**	**100%**	**32.3**

- Normalization: As all the features are binary variables in the same scale of zero and one, no normalization or standardization method is used.
- Features selection (using chi-square contingency and logistic regression): Overall, two sets of features were selected based on the domain knowledge (dropping categories for the nominal/ordinal variables based on the ecological framework, qualified as set 1) or the chi-square selection method available in scikit-learn (qualified as set 2). The final variables for each set were selected using multiple iterations of the logistic regression. Finally, there were different variations of each set depending on the group of classifiers used and the presence of other variables such as other IPV violence or country-level variables.

Choice of classifiers

Different classifiers and models are used to find the best performance by considering the area under the curve (AUC) and the recall. The main metric is the recall (true positive rate) as the classes are not balanced and the interest is in identifying the survivors of IPV who disclosed. AUC and Fscore are used as the best scoring metrics in order to minimize the false negative and false positive rates. The following classifiers and models are used:

- logistic regression,
- support vector machine,
- k-nearest neighbor,
- random forest,
- naives bayes,
- multilayer perceptron, and
- multilevel regression.

These classifiers can be grouped into three main groups: 1) scikit-learn classifiers (logistic regression, support vector machine, k-nearest neighbor, random forest, naives bayes), 2) multilayer perceptron from pytorch, and 3) multilevel regression from lme4 r packages. In addition, three broad types of models were trained: single level, mixed effect (country interaction with other variables), and multilevel level. All the scikit-learn classifiers were trained using the single level and mixed effect models. The MLP from pytorch were trained using the single level only. The only multilevel models used the lme4 r packages. Finally, depending on the types of models, other variables such as other types of IPV (in all the types of models) and country level (only in multilevel regression) variables were included.

Table 4. Summary of the features, classifiers, and models used

Category	**Variables Selected (Chi-square)**	**Scikit-Learn Classifiers**	**Multilayer Perceptron (MLP)**	**Multilevel Regression**
Women's socio-demographic characteristics	Edresp*, EP3*, SourceIncome*, ageyr10* (4/7)	Included in all the single and mixed models (features set 1 and 2)	Included in only the single level models (features set 1 and 2)	Included in only the multi-level models (features set 1 only)
Women's social support	Fam_support (1/1)			
Women's parenting and child abuse	mcv1006, mcv1006a, mcv1006b (3/3)			
Marital traditional norms	earlymarriage*, FCMAR* (2/2)			
Partner's individual characteristics	edpart*, sumdiffage* (2/4)			
Partner's parenting and child abuse	mcv1008, mcv1009 (2/2)			
Partner's other behaviour	Q515R, Q513R, Q516R, men_alcohol_all (4/4)			
Couple relationship dynamics	CONTROLNUM*, Q702R, rQ701d, rQ701b, rQ701a, rQ701c (6/6)			
Women's attitude and gender norms	sQ601c, sQ601d, tQ602d, justify, tQ602c (5//11)			
Other violence	emotvio, econviol (2/2)	Added to the existing single and mixed models (features set 1 and 2)	Added only to the single models (features set 1 and 2)	Added only to the multi-level models (features set 1 only)
Country level variables	country, Island, HDI1*, LAWDV1* (4/6)	Only country is used in these models	Only country is used in these models	All the variables are used
Sets of models created		**4 (2 single, 2 mixed) models for each set**	**2 (single only) models for each set**	**2 (multi-level only)**

*Nominal or ordinal variables.

Data exploration

Through a multivariate logistic regression, the importance of many features is determined. Overall, all of the important features are risk factors except for family support and the couple communication. The main risk factors are from the partner's behavior, women and partner's parenting and child abuse, couple relationship dynamics, some of the traditional marital practices, and other IPV. After multiple iterations of the statistically significant features, the main risk and protective factors remain significant in addition with some women's and partners' demographic characteristics (see tables 5 and 6). Finally, it is important to notice that country is not a significant feature.

Table 5. Final Features Selection

Category	Without Other IPV Variables		With Other IPV Variables		Multilevel Model	
	Set 1 (22)	Set 2 (21)	Set 1 (14)	Set 2 (14)	With Economic (27)	Without Economic (19)
Women's socio-demographic characteristics	EP3_2, EP3_3, ageyr10_2, ageyr10_3, ageyr10_4, ageyr10_5	EP3_2, SourceIncome_2		EP3_2	EP3_2', 'EP3_3', 'SourceIncome_2', 'ageyr10_2', 'ageyr10_3', 'ageyr10_4', 'ageyr10_5'	EP3_2', 'ageyr10_2', 'ageyr10_3'
Women's social support	Fam_support	Fam_support	Fam_support	Fam_support	Fam_support'	Fam_support',
Women's parenting and child abuse	mcv1006, mcv1006a, mcv1006b	mcv1006, mcv1006a, mcv1006b	mcv1006	mcv1006	'mcv1006', 'mcv1006a', 'mcv1006b'	mcv1006'
Marital traditional norms	earlymarriage_1	earlymarriage_0, earlymarriage_1	earlymarriage_1	earlymarriage_1	'earlymarriage_0', 'earlymarriage_1'	earlymarriage_0', 'earlymarriage_1'
Partner's individual characteristics	sumdiffage_1	edpart_1, sumdiffage_1	edpart_1, sumdiffage_1	edpart_1, sumdiffage_1	edpart_1', 'sumdiffage_1'	edpart_1', 'sumdiffage_1'
Partner's parenting and child abuse	mcv1008	mcv1008, mcv1009	mcv1008	mcv1008	mcv1008', 'mcv1009'	mcv1008'
Partner's other behaviour	Q513R, Q515R, men_alcohol_all	Q513R, Q515R, men_alcohol_all	Q513R, Q515R	Q513R, Q515R	Q515R', 'Q513R', 'men_alcohol_all'	'Q515R', 'Q513R'
Couple relationship dynamics	CONTROLNUM_1, CONTROLNUM_2, CONTROLNUM_3, Q702R_2, Q702R_3	CONTROLNUM_1, CONTROLNUM_2, CONTROLNUM_3, Q702R_2, Q702R_3, rQ701d	CONTROLNUM_2, CONTROLNUM_3, Q702R_3	CONTROLNUM_2, CONTROLNUM_3, Q702R_3	'CONTROLNUM_1', 'CONTROLNUM_2', 'CONTROLNUM_3', 'Q702R_2', 'Q702R_3'	'CONTROLNUM_2' , 'CONTROLNUM_3' , 'Q702R_3'
Women's attitude and gender norms	sQ601d		sQ601d		sQ601d'	sQ601d'
Other violence			econviol, emotvio	econviol, emotvio		emotvio', 'econviol'
Country level variables					country'	country'

Table 6. Multivariate Logistic Regression (with and without other IPV)

Variables	witout other IPV		with other IPV	
	Odd Ratio	p-value	Odd ratio	p-value
Intercept	0.033463	0	0.020439	0
emotvio (emotional IPV)	-'	-'	9.241461	0
econviol (economic IPV)	-'	-'	1.466971	0.003
Q515R (partner has other relationship)	1.764911	0	1.364926	0.013
Q513R (partner involves in fight)	1.704207	0	1.468292	0.011
mcv1006 (woman witnessed IPV at home)	1.562833	0	1.468292	0
mcv1006a (Women beaten in childhood)	1.352967	0.004	1.156386	0.22
mcv1006b (Women emotionally abused in childhood)	1.454264	0	1.184594	0.138
mcv1008 (partner witnessed IPV at home)	1.535107	0.001	1.448459	0.012
men_alcohol_all (man drinks alcohol at least once/week)	1.264529	0.012	1.148435	0.189
mcv1009 (Partner was hit as a child)	1.348644	0.007	1.22728	0.102
Fam_support (family support)	0.632737	0	0.684819	0.001
Q516R (Partner has had children with another woman)	0.854619	0.357	0.887098	0.528
rQ701d (partner shares his worries or feelings)	0.725931	0.053	0.863294	0.424
rQ701b (she shares things that happened to her during the day)	0.690044	0.06	0.67848	0.08
rQ701a (he shares things that happened to him during the day)	1.183883	0.369	1.442676	0.083
rQ701c (woman shares her worries or feelings)	1.324983	0.112	1.12221	0.557
sQ601c (A woman's most important role is to take care of her home)	0.995112	0.958	1.009343	0.929
sQ601d (It is natural that men should be the head of the family_	1.371493	0.001	1.360973	0.004
tQ602d (If a woman does not physically fight back, it is not rape)	1.083504	0.475	1.066839	0.605
justify (justify at least one act of violence)	0.91842	0.463	1.071543	0.597
tQ602c (If a woman is raped she has done something careless to put herself in that situation)	1.073045	0.594	1.111933	0.479
CONTROLNUM_1 (one act of controlling behavior) none is reference	1.363834	0.003	1.188153	0.142
CONTROLNUM_2 (2 acts)	2.160414	0	1.469761	0.003
CONTROLNUM_3 (3 acts)	3.453886	0	2.165822	0
Q702R_2 (quarelling sometimes) never is reference	1.431038	0	1.202978	0.06
Q702R_3 (quarelling frequently)	3.391255	0	2.062255	0
earlymarriage_0 (married at 19 or older) not married is reference	1.085456	0.256	1.066732	0.425
earlymarriage_1 (married at 18 or younger)	1.59281	0	1.472262	0
edpart_1 (at most primary level) higher/tech/vocational is reference	1.399059	0.012	1.396683	0.025
edpart_2 (secondary)	1.249071	0.074	1.186609	0.216
sumdiffage_1 (woman is older) Partner at least 9 years older reference	1.387356	0.009	1.54311	0.002
sumdiffage_2 (partner at most 3 years older)	0.955711	0.68	1.025213	0.84
sumdiffage_3 (partner 4 to 8 years older)	0.959541	0.702	1.047179	0.703
edresp_1 (at most primary level) higher/tech/vocational is reference	0.922009	0.56	1.27303	0.122
edresp_2 (secondary)	0.971416	0.789	1.180573	0.17
country_1 (Guyana) Jamaica is reference	1.388605	0.029	1.014098	0.934
country_3 (Trinidad)	1.281332	0.096	1.027882	0.869
country_4 (Suriname)	1.32313	0.068	1.039563	0.822
country_5 (Grenada)	1.113157	0.492	0.978436	0.901
EP3_2 (living with man, not married) currently married is reference	1.557217	0	1.355811	0.032
EP3_3 (regular partner, living apart)	1.757513	0.003	1.466678	0.069
EP3_4 (currently no partner)	1.116948	0.429	1.012072	0.939
FCMAR_0 (non consensual relationship -no) not married is reference	1.32194	0.002	1.240358	0.029
FCMAR_1 (non consensual relationship -yes)	1.30787	0.002	1.266174	0.014
SourceIncome_1 (no income/pension/social services/other) support from partner	1.162067	0.361	1.218109	0.289
SourceIncome_2 (Income from own work)	1.351209	0.013	1.334291	0.033
SourceIncome_4 (Equal share self and partner)	1.137121	0.287	1.183883	0.209
SourceIncome_5 (Support from relatives/friends)	1.320486	0.065	1.451068	0.028
ageyr10_2 (25-34) 15-24 is reference	1.651031	0	1.617853	0.002
ageyr10_3 (35-44)	1.949943	0	1.707278	0.001
ageyr10_4 (45-54)	1.77039	0	1.48127	0.025
ageyr10_5 (55-64)	1.918033	0	1.34851	0.112

B. *Training*

A total of seven classifiers and models are used in the two sets of the features selected. When possible, the models are trained to find the best value for the AUC and Fscore as the class are not balanced. In addition, some key hyperparameters (regularization parameters, criterion, learning rates, and so on) were tuned. The Gridsearch class from the scikit-learn model selection packages was used for most of the classifiers from the scikit-learn group. When possible, class imbalance is controlled to improve the classifiers learning. Finally, the SVM classifier in scikit-learn did not achieve convergence.

Table 7. Summary of the Hyperparameters Tuned during the Training

Classifiers/Models	**Hyperparameter Tuning**	**Scoring Search**
Logistic Regression (LR)	C, solver, learning rate	AUC, Fscore
Support Vector Machine (SVM)	C, solver, learning rate	AUC, Fscore
K-Nearest Neighbors (KNN)*	Number of neighbors, degree of the distance, weights	AUC, Fscore
Random Forest (RF)	Number of decision trees, criterion, maximum of features	AUC, Fscore
Naïve Bayes (NB)*	None	None
Multilayer Perceptron (MLP)	Learning rate	AUC, Fscore
Multilevel Regression*	None	None

*No class imbalance consideration

C. *Post training*

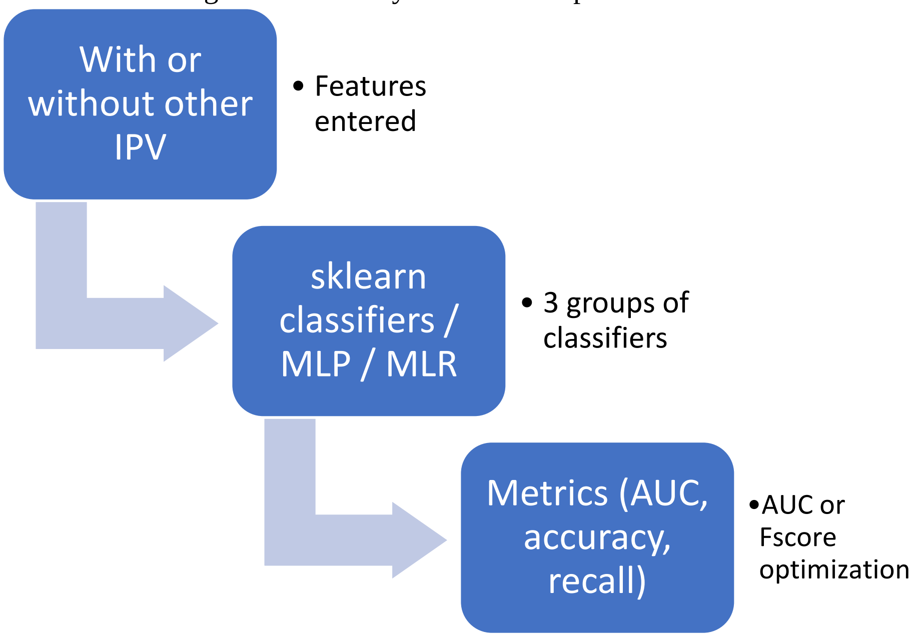
The classical metrics used for classification supervised learning models are used to evaluate the performance of each model. The metrics are:

- Confusion Matrix: Globally informs on the classification accuracy and errors.
- Precision: is the ratio of true positive on the sum of true positive and false positive.
- Recall: is the ration of true positive on the sum of true positive and false negative.
- Fscore: is a combination of precision and recall.
- Accuracy: is the ratio of the number of classes correctly classified.
- Receiver Operating Characteristic (ROC): is a graph that informs on model performance with respect to the true positive rates and false positive rates.
- Area Under the Curve (AUC): is the area that the ROC graphs cover between zero and one. One is the best value.

Results

In this section, the findings from the training are presented in two broad sections: with and without other IPV. For each of the section, the results of the 3 groups of classifiers are presented for the best models' performance considering AUC and Fscore as the optimum metrics. AUC, accuracy, and recall are the metrics considered to select the best models. The results for the mixed models are in annex.

Figure 5. Summary of the results presentation



Without Other Violence

Scikit-learn classifiers

The best three classifiers are logistic regression, random forest, and naïve bayes considering the AUC, the accuracy, and the recall. Logistic regression shows higher AUC or Fscore overall. Even though, it does not have the best score for accuracy and recall as compared to naïve bayes or knn, but it presents the best score for recall-yes, which confirms more consistency in the learning for all the target classes. So, the logistic classifier is the best one. It training as the fscore for the optimum score outperforms just a little the one based on the AUC, considering recall-yes as the main metric. In addition, the learning with set 1 features yields better results. This is a confirmation of the importance of domain knowledge. In summary, the logistic classifier trained through fscore with set 1 features is the best model to consider.

Figure 6. Sklearn Classifiers Summary Results considering AUC as the Optimum Metric

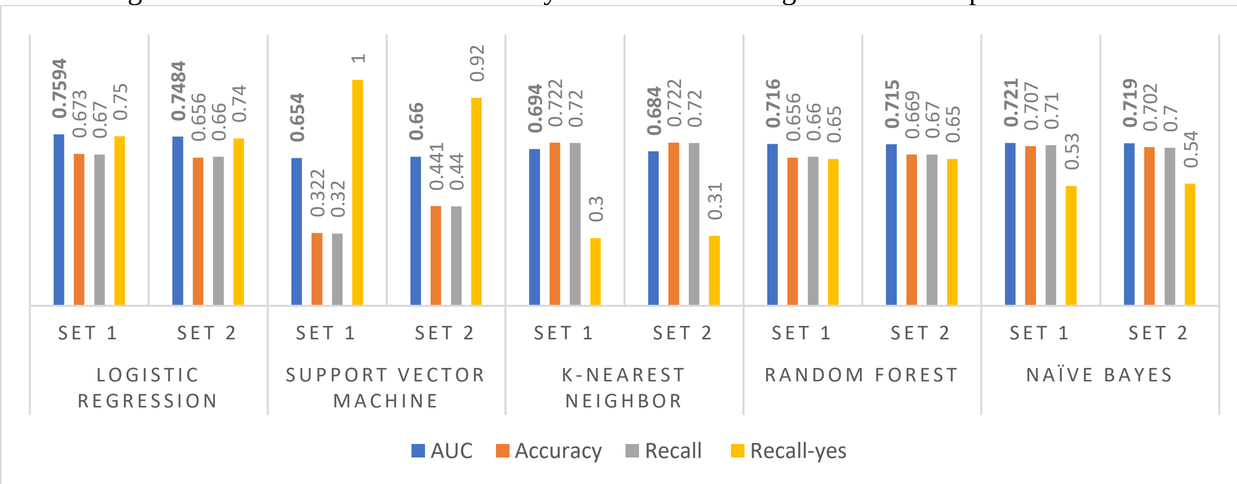
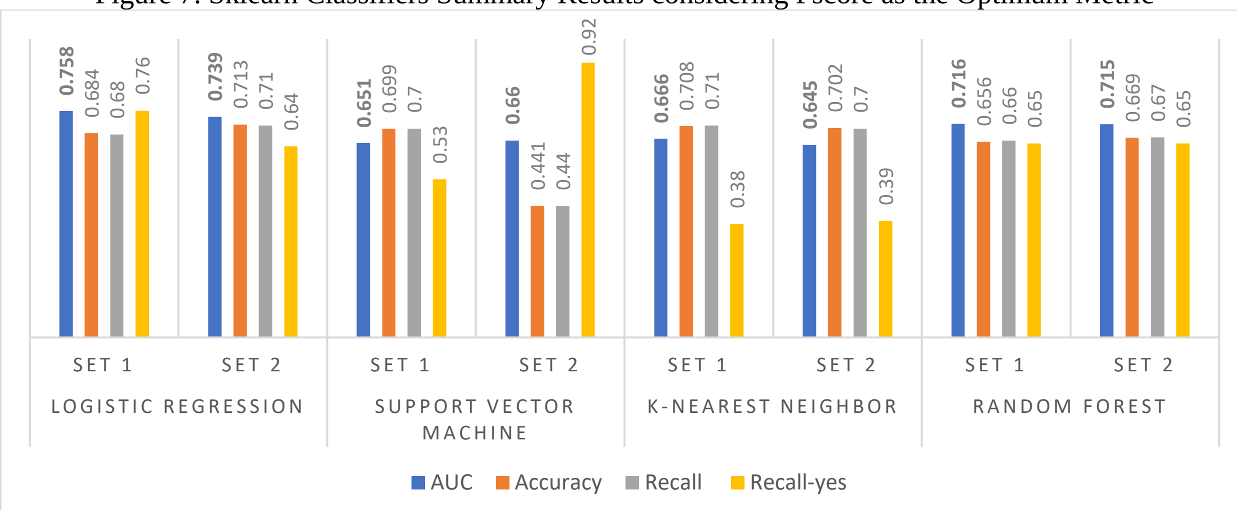


Figure 7. Sklearn Classifiers Summary Results considering Fscore as the Optimum Metric



Multilayer Perceptron

Two sets of networks were trained with respectively two and three layers in each of the sets. The sigmoid is used as the transfer function and these networks can be considered as an extension of the logistic classifier. The results from the three and two layers are close. In such case, the choice of the two layers prevails in order to avoid overfitting. Similar to the logistic classifier, the training based on fscore and set 1 shows the best results for AUC and recall-yes. This network shows better results in terms of accuracy compared to the logistic classifier. However, when recall-yes is considered as the principal metric, the logistic classifier is the best one.

Figure 8. Multilayer Perceptron Summary Results considering AUC as the Optimum Metric

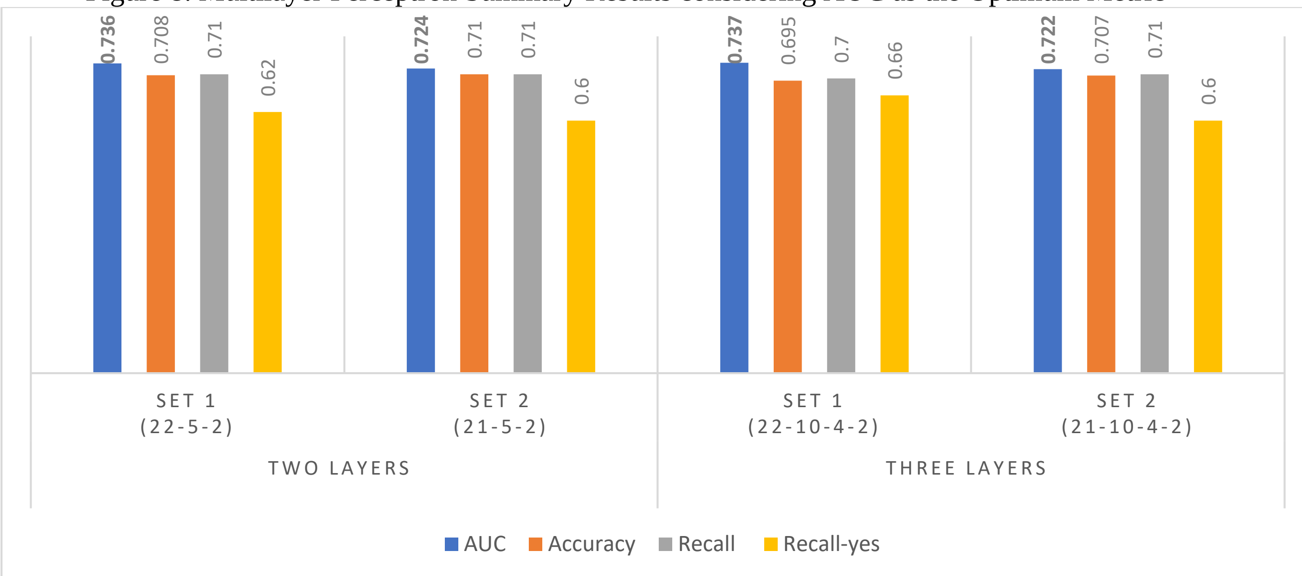
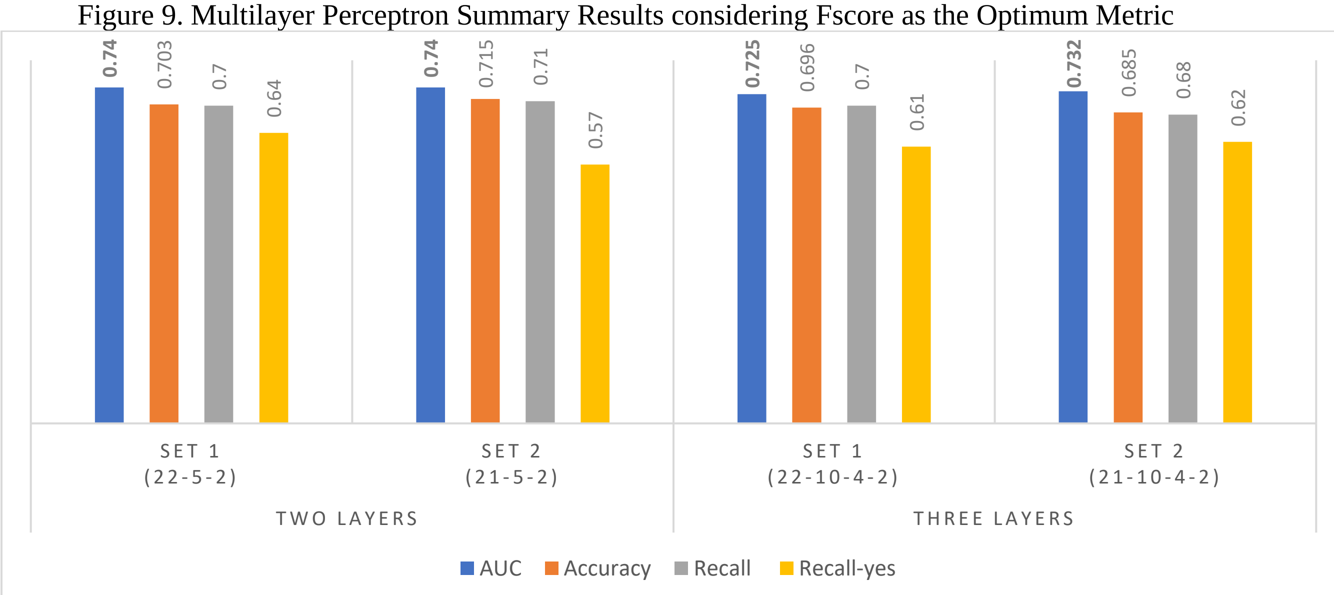


Figure 9. Multilayer Perceptron Summary Results considering Fscore as the Optimum Metric



Multilevel regression

The final multilevel regression model is obtained after a couple of iterations of the models on the significant features. Only 0.02% of the residual variance is explained by the country variation. One can assume that a pooling regression model will yield the same result.

This model includes relatively more features compared to the previous models, but none of the country level variables were kept. It has an AUC of 0.75, closer to the logistic classifier, and the highest accuracy of 0.7536[5], due to class imbalance. Similar to the knn classifier, the model tends to perform well on the class with more datapoints when class imbalance is not controlled for. In that regard, recall will be good for one class and not for the other one, in that case the recall-yes is low (0.42). The logistic classifier still holds the best result.

Table 8. Multilevel Regression without Other IPV

Random effects:					
Groups Name	Variance	Std.Dev.			
country (Intercept)	0.0008064	0.0284			
Number of obs: 4054, groups: country, 5					
Fixed effects:					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.01369	0.19225	-15.676	< 2e-16	***
Q515R	0.53405	0.09392	5.686	1.30e-08	***
Q513R	0.52132	0.13198	3.950	7.81e-05	***
mcv1006	0.45831	0.08836	5.187	2.14e-07	***
mcv1006a	0.29063	0.10469	2.776	0.005500	**
mcv1006b	0.35973	0.10012	3.593	0.000327	***
mcv1008	0.44292	0.12890	3.436	0.000590	***
men_alcohol_all	0.25114	0.09141	2.747	0.006008	**
mcv1009	0.28619	0.10924	2.620	0.008795	**
Fam_support	-0.42819	0.10179	-4.207	2.59e-05	***
sQ601d	0.30973	0.09038	3.427	0.000611	***
CONTROLNUM_1	0.29197	0.10296	2.836	0.004571	**
CONTROLNUM_2	0.77490	0.11219	6.907	4.96e-12	***
CONTROLNUM_3	1.27746	0.11166	11.441	< 2e-16	***
Q702R_2	0.33447	0.08595	3.892	9.96e-05	***
Q702R_3	1.24737	0.12027	10.371	< 2e-16	***
earlymarriage_0	0.28150	0.10246	2.747	0.006005	**
earlymarriage_1	0.69094	0.12848	5.378	7.54e-08	***
edpart_1	0.19729	0.08325	2.370	0.017801	*
sumdiffage_1	0.38353	0.10424	3.679	0.000234	***
EP3_2	0.34569	0.09660	3.579	0.000345	***
EP3_3	0.46136	0.14779	3.122	0.001798	**
SourceIncome_2	0.17658	0.08820	2.002	0.045278	*
ageyr10_2	0.42103	0.13771	3.057	0.002233	**
ageyr10_3	0.56984	0.14168	4.022	5.77e-05	***
ageyr10_4	0.48779	0.15000	3.252	0.001146	**
ageyr10_5	0.59831	0.15949	3.751	0.000176	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

[5] The balanced accuracy is only 0.67.

Table 9. Multilevel Regression Summary Results (without other IPV)

Accuracy : 0.7526	
95% CI : (0.7225, 0.781)	
No Information Rate : 0.6778	
P-Value [Acc > NIR] : 8.428e-07	
Kappa : 0.3691	
Mcnemar's Test P-Value : 1.764e-13	
Sensitivity : 0.4214	
Specificity : 0.9100	
Pos Pred Value : 0.6901	
Neg Pred Value : 0.7679	
Precision : 0.6901	
Recall : 0.4214	
F1 : 0.5233	
Prevalence : 0.3222	
Detection Rate : 0.1358	
Detection Prevalence : 0.1968	
Balanced Accuracy : 0.6657	
'Positive' Class : 1	

Summary of the best models without IPV are included

The best two models so far are the logistic classifier (trained with fscore on the set 1 feature) when considering recall-yes as the principal metric and the 2-layer perceptron (trained with fscore on the set 1 feature) when considering accuracy as the main metric. However, a simple diagnostic of the performance of the two models revealed that the learning tends to achieve a ceiling. The training fscore learning curve for the logistic classifier starts at 0.7 (figure10) or the training for the 2-layer network achieves 0.6 as the highest score (figure 11). In such case, more important features might be necessary. As a reminder, most of the features considered in the two sets are heavily based on the individual and relationship level of the ecological framework. More variables on the community and societal are necessary to improve the performance of the models. Thus, even though the other IPV variables are not included in the ecological framework, they are introduced to test and increase the performance of the models. Another series of training is done with all the classifiers.

Figure 10. Learning Curve for the Logistic Classifier (fscore and set 1)

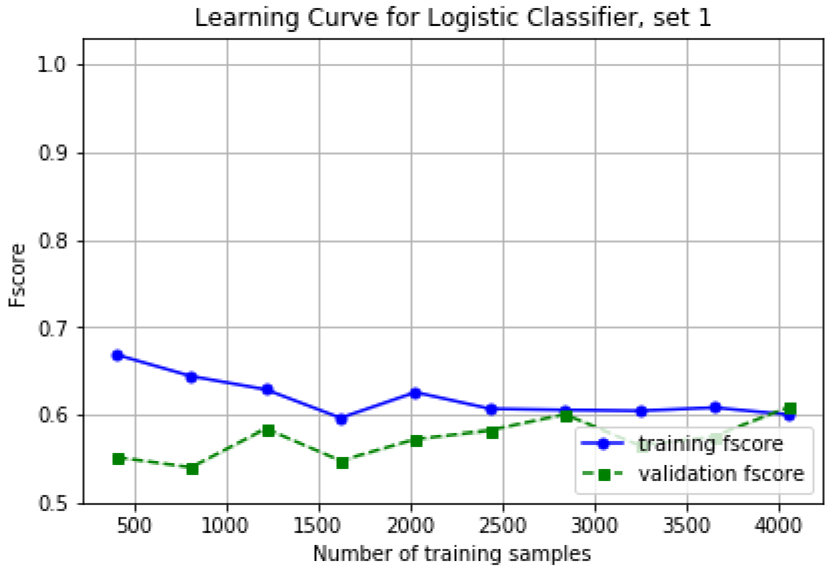
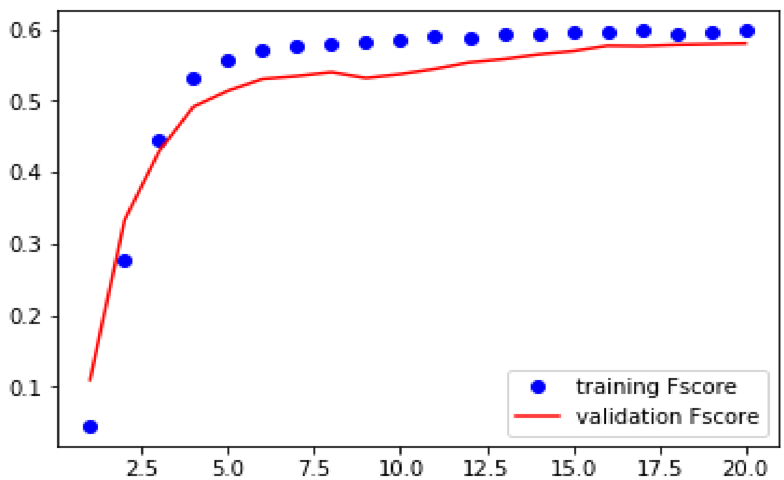


Figure 11. Training Performance of the 2-layer Network (fscore and set 1)



With Other Violence

Scikit-learn classifiers

By including the other IPV variables, the composition of the two sets of features slightly changes. The same trend holds but only the logistic and the random forest classifiers shows the highest performance in terms of AUC, accuracy, and recall. The logistic classifier slightly outperforms the random forest one. The training results considering AUC and fscore as the optimum scoring metrics, the logistic classifiers yield the same performance. Considering the AUC and the recall-yes as the main metrics, the logistic model obtained through AUC is the ultimate choice. The set 1 features still yield the best performance.

Figure 12. Sklearn Classifiers Summary Results considering AUC as the Optimum Metric (other IPV)

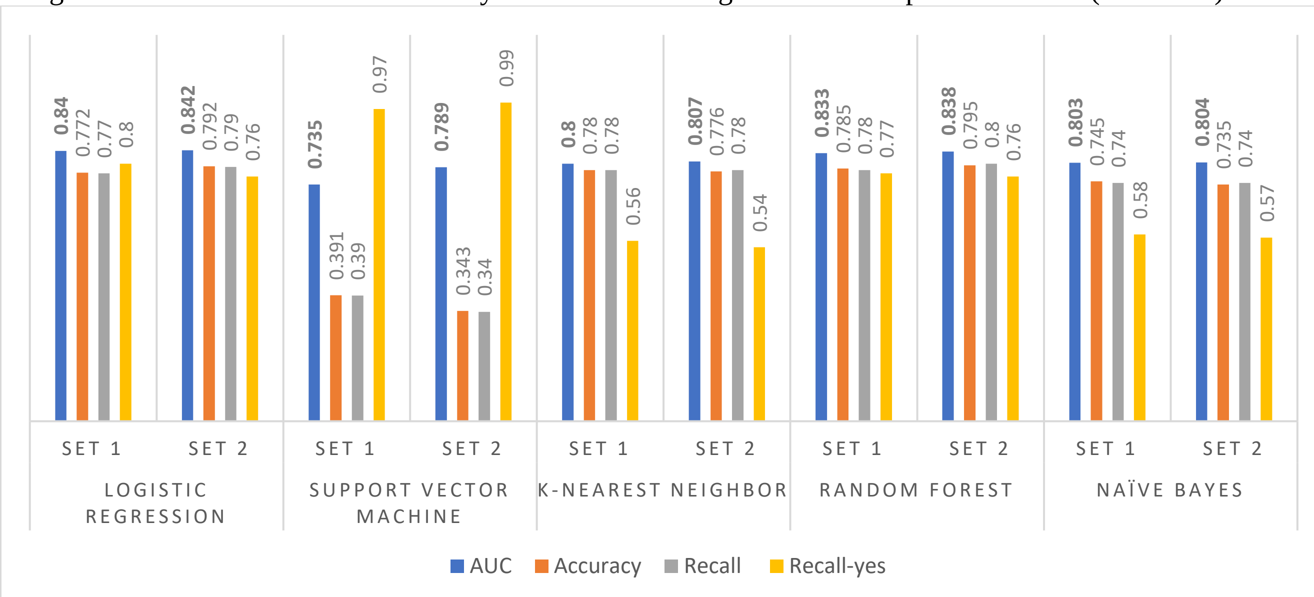
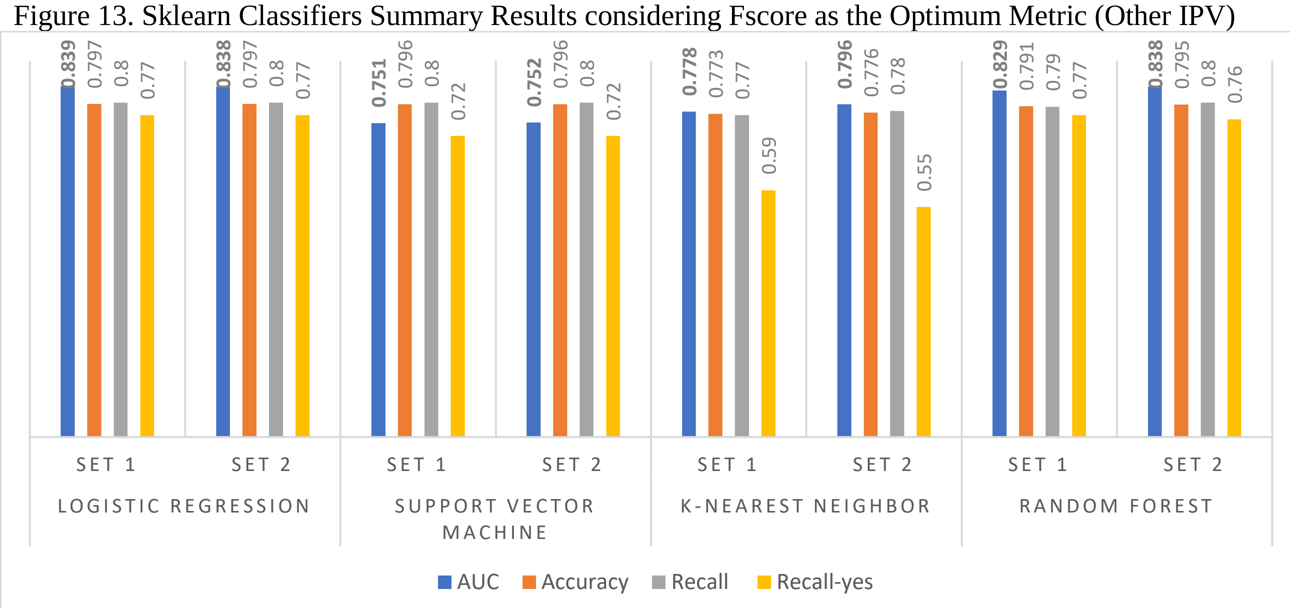


Figure 13. Sklearn Classifiers Summary Results considering Fscore as the Optimum Metric (Other IPV)



Multilayer Perceptron

As in the previous training, the same trends of performance hold between the two- and three-layer network. In such case, the choice of the two-layer is better in order to avoid overfitting. The training performance based on the AUC yield slightly the best performance. Compared to the logistic classifier, the 2-layer network shows slightly better accuracy and recall, but not recall-yes.

Figure 14. Multilayer Perceptron Summary Results considering AUC as the Optimum Metric (Other IPV)

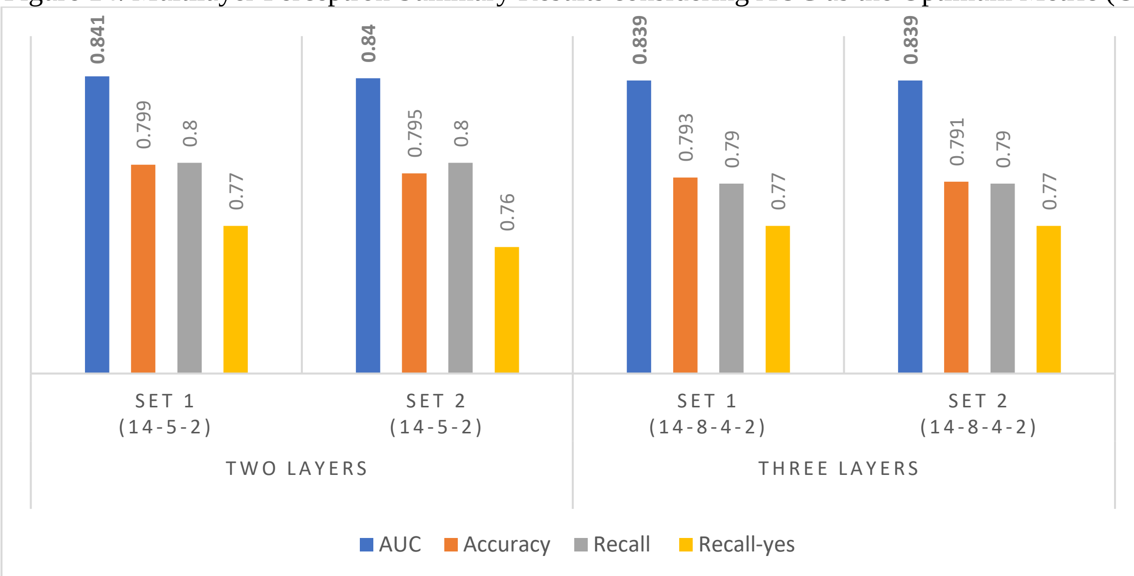
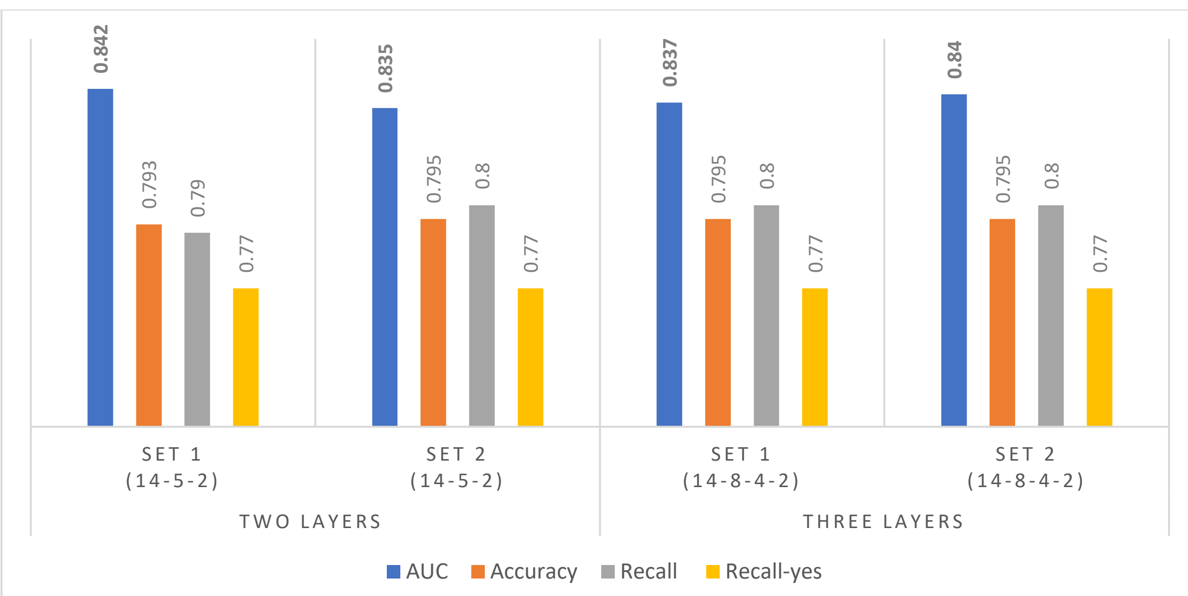


Figure 15. Multilayer Perceptron Summary Results considering Fscore as the Optimum Metric (Other IPV)



Multilevel regression

After a few iterations of the multilevel logistic regression models on the statistically significant features, the final one was obtained. The country variance is null, which suggests that the model is similar to a single level pooling regression. Similar to the previous multilevel logistic regression model, this one does not include any country level variable, but has fewer features. The model has an AUC of 0.835, an accuracy of 0.79, and a recall of 0.61. Both the logistic and the two-layer network classifier have better performance than the multilevel regression model.

Table 10. Multilevel Regression with Other IPV

Random effects:					
Groups	Name	Variance	Std.Dev.		
country	(Intercept)	0	0		
Number of obs: 4054, groups: country, 5					
Fixed effects:					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.86265	0.16905	-16.934	< 2e-16	***
emotvio	2.27380	0.08770	25.928	< 2e-16	***
econviol	0.43881	0.12454	3.523	0.000426	***
Q515R	0.37905	0.10346	3.664	0.000249	***
Q513R	0.42218	0.14697	2.873	0.004072	**
mcv1006	0.45197	0.09635	4.691	2.72e-06	***
mcv1008	0.41046	0.14056	2.920	0.003500	**
Fam_support	-0.40022	0.11076	-3.613	0.000302	***
sQ601d	0.32507	0.09980	3.257	0.001125	**
CONTROLNUM_2	0.37577	0.11745	3.200	0.001376	**
CONTROLNUM_3	0.72779	0.11558	6.297	3.04e-10	***
Q702R_3	0.66214	0.12612	5.250	1.52e-07	***
earlymarriage_0	0.22766	0.10171	2.238	0.025195	*
earlymarriage_1	0.56867	0.12931	4.398	1.09e-05	***
edpart_1	0.26317	0.09116	2.887	0.003890	**
sumdiffage_1	0.45983	0.11360	4.048	5.17e-05	***
EP3_2	0.21234	0.10296	2.062	0.039169	*
ageyr10_2	0.21374	0.10767	1.985	0.047135	*
ageyr10_3	0.23571	0.10688	2.205	0.027425	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Table 11. Multilevel Regression Summary Results (with other IPV)

Accuracy : 0.7906	
95% CI : (0.762, 0.8172)	
No Information Rate : 0.6778	
P-Value [Acc > NIR] : 9.999e-14	
Kappa : 0.5057	
Mcnemar's Test P-Value : 0.02157	
Sensitivity : 0.6179	
Specificity : 0.8727	
Pos Pred Value : 0.6976	
Neg Pred Value : 0.8277	
Precision : 0.6976	
Recall : 0.6179	
F1 : 0.6553	
Prevalence : 0.3222	
Detection Rate : 0.1991	
Detection Prevalence : 0.2854	
Balanced Accuracy : 0.7453	
'Positive' Class : 1	

Summary of the best models when other IPV are included

With the inclusion of other IPV, the performance of all the models slightly improve, and the same classifiers (logistic and 2-layer network) hold the best results (when trained with AUC on the set 1 feature). The diagnostic of these models still revealed that the fscore has improved to 0.7 and other important features need to added.

Figure 16. Learning Curve for the Logistic Classifier (other, AUC, and set 1)

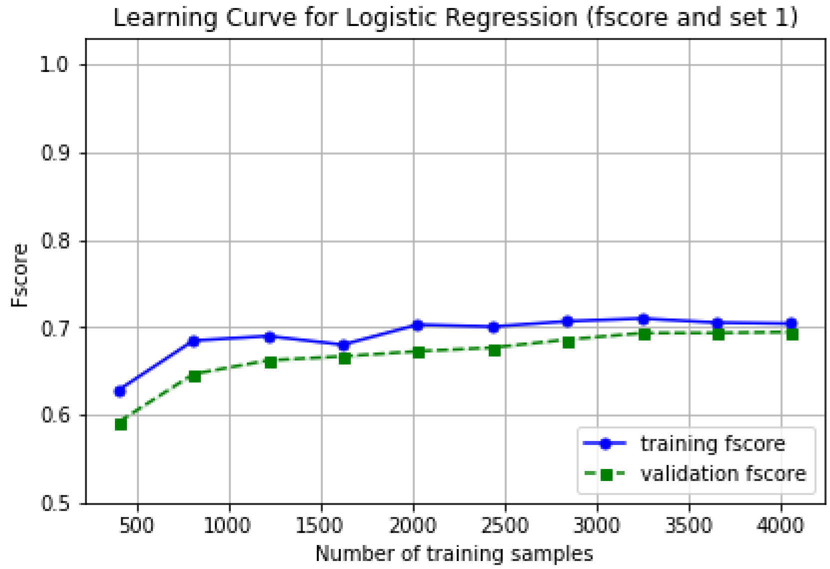
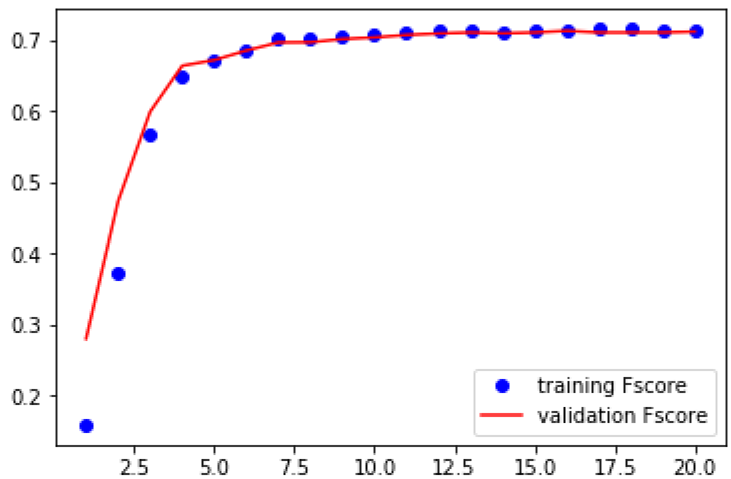


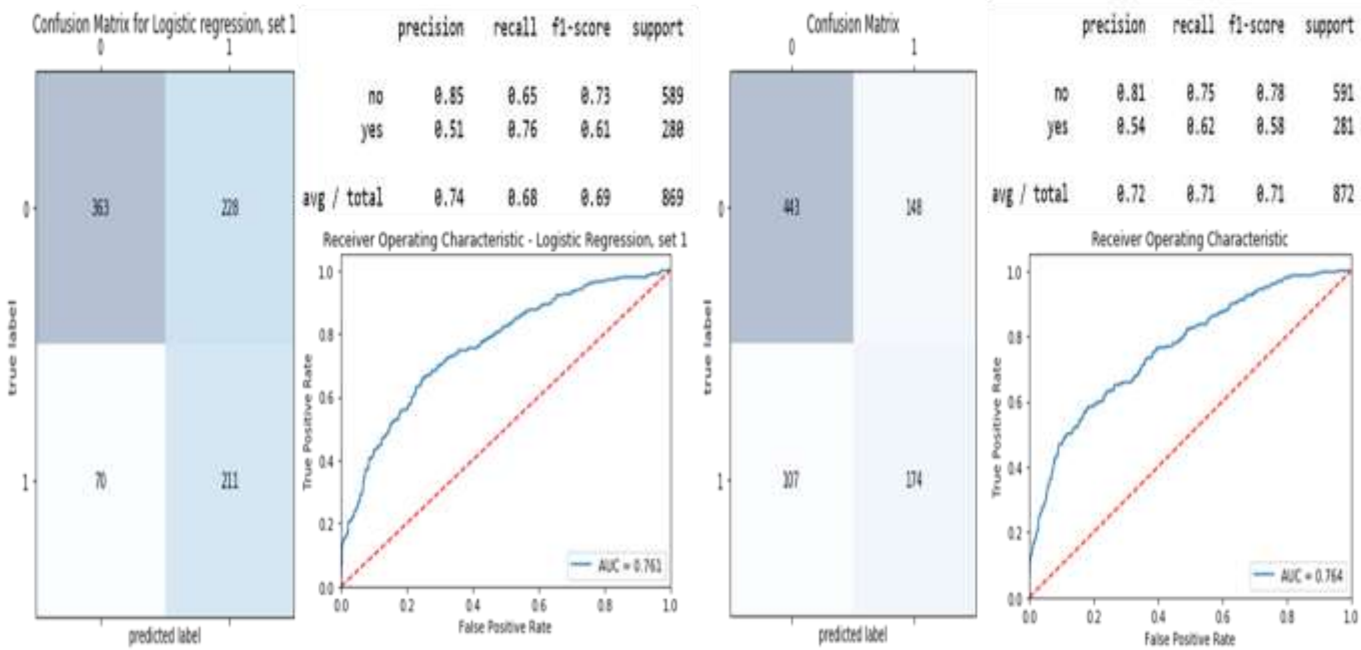
Figure 17. Training Performance of the 2-layer Network (other, AUC, and set 1)



Testing / Model Evaluation

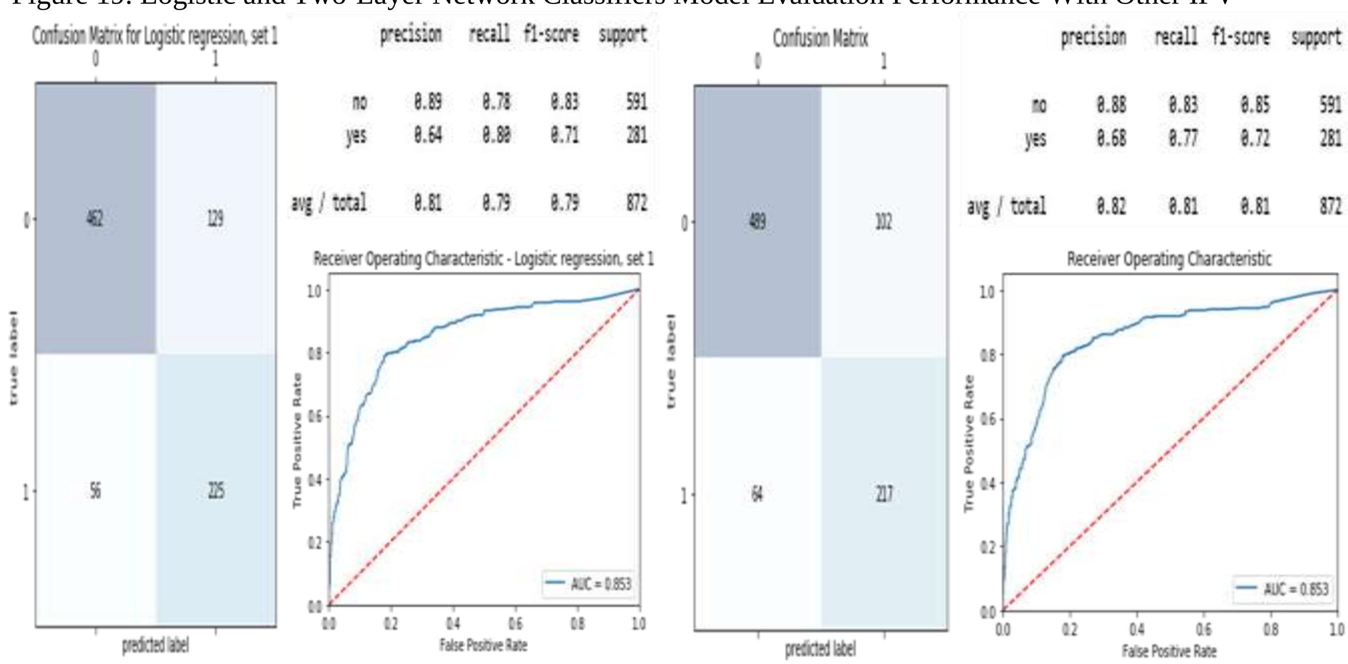
The best two classifiers are logistic and two-layer network whether or not the other IPV variables are included. As expected, their evaluation yields the same result as in the training. The models slightly improve when the other IPV variables are included, and the classifiers have the same performance. The most important features are related to the couple relationship dynamics, the partners' characteristics, the women and partners' experience as a child, some of the traditional marital norms (early marriage), and the family support, which is the only protective factor. However, even with the inclusion of other IPV, more important features need to be considered.

Figure 18. Logistic and Two-Layer Network Classifiers Model Evaluation Performance Without Other IPV*



*Accuracy: Logistic (0.658), 2-layer network (0.708)

Figure 19. Logistic and Two-Layer Network Classifiers Model Evaluation Performance With Other IPV*



*Accuracy: Logistic (0.788), 2-layer network (0.810)

Conclusion

The interest in using machine learning techniques to predict physical and/or sexual IPV can be capital by informing both preventive interventions on the main features/risk factors and responsive actions in being able to better identify potential survivors. The project used the ecological framework as the main theoretical foundations to build the features. Most of the variables are at the individual and relationships level. The performance of the models can be improved by adding other features such as other IPV. The need to add other features from the community or societal levels is vital to ensure a complete learning of the models. Other considerations are necessary in future research:

- Feature scaling: One can normalize or standardize the features to test any difference in terms of performance.
- Inclusion of interaction: It is important to test the interaction between many features. The random forest results suggest the need to investigate these potential interactions.
- Hierarchical modeling: As the ecological framework suggests, the risk and preventive factors are organized hierarchically. It is vital to consider this structure in defining the models.

Finally, as IPV is self-reported, the performance obtained through these models might be the best ones one can achieve.

Annex:

Table A. 1. Summary of the Variables Classified by the Ecological Framework Levels

Variable	Label	Category
sexphys	Lifetime physical and/or sexual violence	Target
w_religion	Religion	Women's socio-demographic characteristics[a]
ageyr10	Respondent age (10 year)	
EP3	Current partnership status	
edresp	Educational attaintment	
employstatus	Main activities during past week	
SourceIncome	Main source of income	
ETHNICITY	Ethnicity	
Fam_support	Family support	Women's social support[b]
mcv1006	Her mother was hit by mother's husband	Women's parenting and child abuse[a]
mcv1006a	Women beaten in childhood	
mcv1006b	She was insulted or humilated as a child	
earlymarriage	Age at first union	Marital traditional norms[d]
FCMAR	Non consensual relationship	
p_employ	Partner is working	Partner's individual characteristics[b]
agepartner	Partner's age	
edpart	Partner's education	
sumdiffage	Age difference	
mcv1008	Partner's mother was hit by mother's husband	Partner's parenting and child abuse[b]
mcv1009	Partner was hit as a child	
men_alcohol_all	Partner drinks alcohol at least once a week	Partner's other behaviour[b]
Q513R	Partner has been involved in a physical fight with another man	
Q515R	Partner has had another relationship	
Q516R	Partner has had children with another woman	
Q702R	Frequency of quarrelling among couple	Couple relationship dynamics[b]
rQ701a	Things that have happened to him in the day	
rQ701b	Things that have happened to her during the day	
rQ701c	Her worries or feelings	
rQ701d	His worries or feelings	
CONTROLNUM	Number of acts of controlling behaviors	
Justify	Justify at least one act of IPV	Women's attitude and gender norms[c]
sQ601a	It is wife's obligation to have sex with husband	
sQ601b	Women and men should share authority in the family	
sQ601c	A woman's most important role is to take care of her home	
sQ601d	It is natural that men should be the head of the family	
sQ601e	A wife should always obey her husband even if she disagrees	

sQ601f	A woman should be able to spend her own money	
tQ602a	Violence between husband and wife is a private matter	
tQ602b	A woman should tolerate violence to keep her family together	
tQ602c	If a woman is raped she has done something careless to put herself in that situation	
tQ602d	If a woman does not physically fight back, it is not rape	
econviol	Lifetime economic violence	Other violence[x]
emotvio	Lifetime emotional violence	
HighIncome	High Income Country	Country level variables[y]
Island	Island	
Governance1	Voice and Accountability percentile	
HDI1	Human Development Index Ranking	
LAWDV1	rule of law percentile	
country	Country	Nominal[y]

a: individual, **b**: relationships, **c**: community, **d**: societal, **x**: other variables, **y**: country level variables

Table A. 2. Single level model training using the AUC as optimum metric

Classifiers	Set	Hyperparameters		Precision	Recall	Fscore	Accuracy	AUC
Logistic Regression	set 1	C=1000, solver=saga, lr=1	no	0.84	0.63	0.72	0.673	0.7594
			yes	0.5	0.75	0.6		
			total	0.73	0.67	0.68		
	set 2	C=1, solver=saga, lr=1	no	0.83	0.61	0.71	0.656	0.7484
			yes	0.48	0.74	0.58		
			total	0.72	0.66	0.67		
Support Vector Machine	set 1	C=100, kernel=sigmoid, gamma=0.1, coefficient=2	no	0	0	0	0.322	0.654
			yes	0.32	1	0.49		
			total	0.1	0.32	0.16		
	set 2	C=10, kernel=rbf, gamma=1	no	0.85	0.21	0.34	0.441	0.66
			yes	0.36	0.92	0.51		
			total	0.69	0.44	0.4		
K-Nearest Neighbor	set 1	Number of neighbors=7, degree=7, weights=uniform	no	0.74	0.92	0.82	0.722	0.694
			yes	0.64	0.3	0.41		
			total	0.71	0.72	0.69		
	set 2	Number of neighbors=7, degree=7, weights=uniform	no	0.74	0.92	0.82	0.722	0.684
			yes	0.64	0.31	0.42		
			total	0.71	0.72	0.69		
Random Forest	set 1	criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 15	no	0.8	0.66	0.72	0.656	0.716
			yes	0.48	0.65	0.55		
			total	0.69	0.66	0.67		
	set 2	criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 15	no	0.8	0.68	0.74	0.669	0.715
			yes	0.49	0.65	0.56		
			total	0.7	0.67	0.68		
Naïve Bayes	set 1	none	no	0.78	0.79	0.79	0.707	0.721
			yes	0.55	0.53	0.54		
			total	0.7	0.71	0.71		
	set 2	none	no	0.78	0.78	0.78	0.702	0.719
			yes	0.54	0.54	0.54		
			total	0.7	0.7	0.7		

Table A. 3. Single level model training using the Fscore as optimum metric

Classifiers	Set	Hyperparameters		Precision	Recall	Fscore	Accuracy	AUC
Logistic Regression	set 1	C': 1.0, 'solver': 'saga', 'tol': 1.0	no	0.85	0.65	0.73	0.684	0.758
			yes	0.51	0.76	0.61		
			total	0.74	0.68	0.69		
	set 2	C': 0.01, 'solver': 'saga', 'tol': 0.1	no	0.81	0.75	0.78	0.713	0.739
			yes	0.55	0.64	0.59		
			total	0.73	0.71	0.72		
Support Vector Machine	set 1	C': 100.0, 'coef0': -1, 'degree': 3, 'gamma': 0.01, 'kernel': 'poly', 'tol': 1e-05	no	0.78	0.78	0.78	0.699	0.651
			yes	0.53	0.53	0.53		
			total	0.7	0.7	0.7		
	set 2	C': 10.0, 'gamma': 1.0, 'kernel': 'rbf', 'tol': 1e-05	no	0.85	0.21	0.34	0.441	0.66
			yes	0.36	0.92	0.51		
			total	0.69	0.44	0.4		
K-Nearest Neighbor	set 1	n_neighbors': 3, 'p': 2, 'weights': 'uniform'	no	0.75	0.86	0.8	0.708	0.666
			yes	0.57	0.38	0.46		
			total	0.69	0.71	0.69		
	set 2	n_neighbors': 3, 'p': 2, 'weights': 'distance'	no	0.74	0.85	0.79	0.702	0.645
			yes	0.55	0.39	0.45		
			total	0.68	0.7	0.69		
Random Forest	set 1	criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 15	no	0.8	0.66	0.72	0.656	0.716
			yes	0.48	0.65	0.55		
			total	0.69	0.66	0.67		
	set 2	criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 15	no	0.8	0.68	0.74	0.669	0.715
			yes	0.49	0.65	0.56		
			total	0.7	0.67	0.68		

Table A. 4. Single level model training using the AUC as optimum metric (including the other IPV)

Classifiers	Set	Hyperparameters		Precision	Recall	Fscore	Accuracy	AUC
Logistic Regression	set 1	C': 0.001, 'solver': 'liblinear', 'tol': 0.1	no	0.89	0.76	0.82	0.772	0.84
			yes	0.61	0.8	0.69		
			total	0.8	0.77	0.78		
	set 2	C': 0.001, 'solver': 'sag', 'tol': 0.01	no	0.88	0.8	0.84	0.792	0.842
			yes	0.65	0.76	0.7		
			total	0.8	0.79	0.8		
Support Vector Machine	set 1	C': 100.0, 'gamma': 1.0, 'kernel': 'rbf', 'tol': 1e-05	no	0.89	0.12	0.2	0.391	0.735
			yes	0.34	0.97	0.51		
			total	0.72	0.39	0.3		
	set 2	C': 10.0, 'coef0': 1, 'gamma': 0.1, 'kernel': 'sigmoid', 'tol': 1e-05	no	0.88	0.04	0.07	0.343	0.789
			yes	0.33	0.99	0.49		
			total	0.7	0.34	0.21		
K-Nearest Neighbor	set 1	n_neighbors': 7, 'p': 2, 'weights': 'uniform	no	0.81	0.89	0.85	0.78	0.8
			yes	0.7	0.56	0.62		
			total	0.77	0.78	0.77		
	set 2	n_neighbors': 7, 'p': 2, 'weights': 'uniform'	no	0.8	0.89	0.84	0.776	0.807
			yes	0.69	0.54	0.61		
			total	0.77	0.78	0.77		
Random Forest	set 1	criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 15	no	0.88	0.79	0.83	0.785	0.833
			yes	0.64	0.77	0.7		
			total	0.8	0.78	0.79		
	set 2	criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 15	no	0.88	0.81	0.84	0.795	0.838
			yes	0.66	0.76	0.71		
			total	0.81	0.8	0.8		
Naïve Bayes	set 1	none	no	0.8	0.82	0.81	0.745	0.803
			yes	0.61	0.58	0.59		
			total	0.74	0.74	0.74		
	set 2	none	no	0.8	0.81	0.81	0.735	0.804
			yes	0.59	0.57	0.58		
			total	0.73	0.74	0.73		

Table A. 5. Single level model training using the Fscore as optimum metric (including the other IPV)

Classifiers	**Set**	**Hyperparameters**		**Precision**	**Recall**	**Fscore**	**Accuracy**	**AUC**
Logistic Regression	set 1	C': 0.01, solver': 'liblinear', 'tol': 1e-05	no	0.88	0.81	0.84	0.797	0.839
			yes	0.66	0.77	0.71		
			total	0.81	0.8	0.8		
	set 2	C': 0.01, solver': 'liblinear', tol': 1e-05	no	0.88	0.81	0.84	0.797	0.838
			yes	0.66	0.77	0.71		
			total	0.81	0.8	0.8		
Support Vector Machine	set 1		no	0.86	0.83	0.85	0.796	0.751
			yes	0.67	0.72	0.7		
			total	0.8	0.8	0.8		
	set 2	C': 1000.0, 'coef0': 1, 'gamma': 0.01, 'kernel': 'sigmoid', 'tol': 1e-05	no	0.86	0.83	0.85	0.796	0.752
			yes	0.67	0.72	0.7		
			total	0.8	0.8	0.8		
K-Nearest Neighbor	set 1	n_neighbors': 5, 'p': 2, 'weights': 'uniform'	no	0.81	0.86	0.84	0.773	0.778
			yes	0.67	0.59	0.62		
			total	0.77	0.77	0.77		
	set 2	n_neighbors': 5, 'p': 2, 'weights': 'uniform'	no	0.8	0.88	0.84	0.776	0.796
			yes	0.69	0.55	0.61		
			total	0.77	0.78	0.77		
Random Forest	set 1	criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 10	no	0.88	0.8	0.84	0.791	0.829
			yes	0.65	0.77	0.7		
			total	0.8	0.79	0.79		
	set 2	criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 15	no	0.88	0.81	0.84	0.795	0.838
			yes	0.66	0.76	0.71		
			total	0.81	0.8	0.8		

Table A. 6. Country-effect model training using the AUC as optimum metric

Classifiers	Set	Hyperparameters		Precision	Recall	Fscore	Accuracy	AUC
Logistic Regression	set 1	C': 1.0, 'solver': 'liblinear', 'tol': 1.0	no	0.8	0.77	0.78	0.71	0.729
			yes	0.55	0.59	0.57		
			total	0.72	0.71	0.71		
	set 2	C': 0.01, 'solver': 'sag', 'tol': 0.1	no	0.79	0.76	0.78	0.705	0.737
			yes	0.54	0.59	0.56		
			total	0.71	0.71	0.71		
Support Vector Machine	set 1	C': 1000.0, 'coef0': -2, 'gamma': 1.0, 'kernel': 'sigmoid', 'tol': 1e-05	no	0.76	0.88	0.82	0.728	0.3
			yes	0.72	0.4	0.49		
			total	0.71	0.73	0.71		
	set 2	C': 10.0, 'coef0': 1, 'degree': 2, 'gamma': 1e-05, 'kernel': 'poly', 'tol': 1e-05	no	0	0	0	0.322	0.406
			yes	0.32	1	0.49		
			total	0.1	0.32	0.16		
K-Nearest Neighbor	set 1	n_neighbors': 7, 'p': 2, 'weights': 'uniform'	no	0.75	0.81	0.78	0.688	0.696
			yes	0.52	0.42	0.47		
			total	0.67	0.69	0.68		
	set 2	n_neighbors': 5, 'p': 2, 'weights': 'uniform'	no	0.75	0.84	0.79	0.705	0.675
			yes	0.56	0.42	0.48		
			total	0.69	0.71	0.68		
Random Forest	set 1	criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 15	no	0.8	0.73	0.76	0.693	0.726
			yes	0.52	0.61	0.56		
			total	0.71	0.69	0.7		
	set 2	criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 15	no	0.81	0.7	0.75	0.685	0.727
			yes	0.51	0.65	0.57		
			total	0.71	0.68	0.69		
Naïve Bayes	set 1	none	no	0.77	0.82	0.79	0.71	0.701
			yes	0.56	0.47	0.51		
			total	0.7	0.71	0.7		
	set 2	none	no	0.77	0.81	0.79	0.708	0.723
			yes	0.55	0.49	0.52		
			total	0.7	0.71	0.7		

Table A. 7. Country-effect model training using the Fscore as optimum metric

Classifiers	Set	Hyperparameters		Precision	Recall	Fscore	Accuracy	AUC
Logistic Regression	set 1	C': 0.1, 'solver': 'liblinear', 'tol': 1e-05	no	0.8	0.79	0.8	0.725	0.726
			yes	0.57	0.59	0.58		
			total	0.73	0.72	0.73		
	set 2	C': 10.0, solver': 'liblinear', tol': 0.1	no	0.8	0.76	0.78	0.71	0.735
			yes	0.54	0.61	0.58		
			total	0.72	0.71	0.71		
Support Vector Machine	set 1	C': 10.0, 'coef0': -2, 'gamma': 1.0, 'kernel': 'sigmoid', 'tol': 1e-05	no	0.76	0.78	0.77	0.688	0.686
			yes	0.52	0.49	0.5		
			total	0.68	0.69	0.69		
	set 2		no					
			yes					
			total					
K-Nearest Neighbor	set 1	n_neighbors': 3, 'p': 2, 'weights': 'uniform'	no	0.75	0.77	0.76	0.67	0.617
			yes	0.49	0.47	0.48		
			total	0.67	0.67	0.67		
	set 2	n_neighbors': 5, 'p': 2, 'weights': 'uniform'	no	0.75	0.84	0.79	0.705	0.675
			yes	0.56	0.42	0.48		
			total	0.69	0.71	0.69		
Random Forest	set 1	criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 15	no	0.8	0.73	0.76	0.693	0.726
			yes	0.52	0.61	0.56		
			total	0.71	0.69	0.7		
	set 2	criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 15	no	0.81	0.7	0.75	0.685	0.727
			yes	0.51	0.65	0.57		
			total	0.71	0.68	0.69		

Table A. 8. Country-effect model training using the AUC as optimum metric (including other IPV)

Classifiers	Set	Hyperparameters		Precision	Recall	Fscore	Accuracy	AUC
Logistic Regression	set 1	C': 0.1, 'solver': 'sag', 'tol': 1.0	no	0.88	0.81	0.84	0.796	0.822
			yes	0.66	0.77	0.71		
			total	0.81	0.8	0.8		
	set 2	C': 100.0, 'solver': 'saga', 'tol': 0.1	no	0.87	0.82	0.84	0.793	0.827
			yes	0.66	0.74	0.7		
			total	0.8	0.79	0.8		
Support Vector Machine	set 1	C': 1000.0, 'coef0': -1, 'degree': 1, 'gamma': 1e-05, 'kernel': 'poly', 'tol': 1e-05	no	0	0	0	0.322	0.807
			yes	0.32	1	0.49		
			total	0.1	0.32	0.16		
	set 2	C': 10.0, 'coef0': 1, 'degree': 4, 'gamma': 0.1, 'kernel': 'poly', 'tol': 1e-05	no	0	0	0	0.322	0.758
			yes	0.32	1	0.49		
			total	0.1	0.32	0.16		
K-Nearest Neighbor	set 1	n_neighbors': 7, 'p': 2, 'weights': 'uniform'	no	0.83	0.86	0.84	0.78	0.792
			yes	0.67	0.62	0.64		
			total	0.78	0.78	0.78		
	set 2	n_neighbors': 7, 'p': 2, 'weights': 'uniform'	no	0.82	0.88	0.85	0.789	0.798
			yes	0.71	0.59	0.64		
			total	0.78	0.79	0.78		
Random Forest	set 1	criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 10	no	0.87	0.81	0.84	0.714	0.813
			yes	0.66	0.75	0.7		
			total	0.8	0.79	0.8		
	set 2	criterion': 'entropy', 'max_features': 'log2', 'n_estimators': 15	no	0.88	0.78	0.83	0.778	0.818
			yes	0.63	0.77	0.69		
			total	0.8	0.78	0.78		
Naïve Bayes	set 1	none	no	0.77	0.82	0.79	0.71	0.701
			yes	0.56	0.47	0.51		
			total	0.7	0.71	0.7		
	set 2	none	no	0.77	0.81	0.79	0.708	0.723
			yes	0.55	0.49	0.52		
			total	0.7	0.71	0.7		

Table A. 9. Country-effect model training using the Fscore as optimum metric (including other IPV)

Classifiers	**Set**	**Hyperparameters**		**Precision**	**Recall**	**Fscore**	**Accuracy**	**AUC**
Logistic Regression	set 1	C': 0.1, 'solver': 'sag', 'tol': 1.0	no	0.88	0.81	0.84	0.796	0.822
			yes	0.66	0.77	0.71		
			total	0.81	0.8	0.8		
	set 2	C': 0.001, 'solver': 'sag', 'tol': 1e-05	no	0.87	0.82	0.84	0.795	0.825
			yes	0.66	0.74	0.7		
			total	0.8	0.8	0.8		
Support Vector Machine	set 1	C': 0.1, 'coef0': -1, 'degree': 4, 'gamma': 0.1, 'kernel': 'poly', 'tol': 1e-05	no	0.78	0.04	0.07	0.339	0.607
			yes	0.33	0.98	0.49		
			total	0.63	0.34	0.2		
	set 2	C': 100.0, 'coef0': -1, 'degree': 4, 'gamma': 0.1, 'kernel': 'poly', 'tol': 1e-05	no	1	0	0.01	0.325	0.672
			yes	0.32	1	0.49		
			total	0.78	0.32	0.16		
K-Nearest Neighbor	set 1	n_neighbors': 7, 'p': 2, 'weights': 'distance'	no	0.83	0.86	0.84	0.783	0.785
			yes	0.67	0.63	0.65		
			total	0.78	0.78	0.78		
	set 2	n_neighbors': 7, 'p': 2, 'weights': 'uniform'	no	0.82	0.88	0.85	0.789	0.798
			yes	0.71	0.59	0.64		
			total	0.78	0.79	0.78		
Random Forest	set 1	criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 15	no	0.88	0.81	0.84	0.795	0.809
			yes	0.66	0.76	0.7		
			total	0.81	0.8	0.8		
	set 2	criterion': 'gini', 'max_features': 'log2', 'n_estimators': 5	no	0.87	0.82	0.84	0.794	0.816
			yes	0.66	0.74	0.7		
			total	0.8	0.79	0.8		

Table A. 10. Multilayer Perceptron training

Network	Set		Precision	Recall	Fscore	Accuracy	AUC
Scoring search = AUC							
Two layers	set 1 (22-5-2)	no	0.81	0.75	0.78	0.708	0.736
		yes	0.54	0.62	0.58		
		total	0.72	0.71	0.71		
	set 2 (21-5-2)	no	0.8	0.76	0.78	0.71	0.724
		yes	0.55	0.6	0.57		
		total	0.72	0.71	0.71		
Three layers	set 1 (22-10-4-2)	no	0.82	0.71	0.76	0.695	0.737
		yes	0.52	0.66	0.58		
		total	0.72	0.7	0.7		
	set 2 (21-10-4-2)	no	0.8	0.76	0.78	0.707	0.722
		yes	0.54	0.6	0.57		
		total	0.72	0.71	0.71		
Scoring search = Fscore							
Two layers	set 1 (22-5-2)	no	0.81	0.74	0.77	0.703	0.74
		yes	0.53	0.64	0.58		
		total	0.72	0.7	0.71		
	set 2 (21-5-2)	no	0.8	0.79	0.79	0.715	0.74
		yes	0.55	0.57	0.57		
		total	0.72	0.71	0.72		
Three layers	set 1 (22-10-4-2)	no	0.8	0.74	0.77	0.696	0.725
		yes	0.52	0.61	0.57		
		total	0.71	0.7	0.7		
	set 2 (21-10-4-2)	no	0.8	0.71	0.75	0.685	0.732
		yes	0.51	0.62	0.56		
		total	0.71	0.68	0.69		

Table A. 11. Multilayer Perceptron training (including other IPV)

Network	Set		Precision	Recall	Fscore	Accuracy	AUC
Scoring search = AUC							
Two layers	set 1 (14-5-2)	no	0.88	0.81	0.85	0.799	0.841
		yes	0.66	0.77	0.71		
		total	0.8	0.8	0.8		
	set 2 (14-5-2)	no	0.88	0.81	0.84	0.795	0.84
		yes	0.66	0.76	0.71		
		total	0.81	0.8	0.8		
Three layers	set 1 (14-8-4-2)	no	0.88	0.8	0.84	0.793	0.839
		yes	0.65	0.77	0.71		
		total	0.81	0.79	0.8		
	set 2 (14-8-4-2)	no	0.88	0.8	0.84	0.791	0.839
		yes	0.65	0.77	0.7		
		total	0.81	0.79	0.79		
Scoring search = Fscore							
Two layers	set 1 (14-5-2)	no	0.88	0.8	0.84	0.793	0.842
		yes	0.65	0.77	0.7		
		total	0.81	0.79	0.8		
	set 2 (14-5-2)	no	0.88	0.81	0.84	0.795	0.835
		yes	0.66	0.77	0.71		
		total	0.81	0.8	0.8		
Three layers	set 1 (14-8-4-2)	no	0.88	0.81	0.84	0.795	0.837
		yes	0.66	0.77	0.71		
		total	0.81	0.8	0.8		
	set 2 (14-8-4-2)	no	0.88	0.81	0.84	0.795	0.84
		yes	0.65	0.77	0.71		
		total	0.81	0.8	0.8		