

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

**Capstone Project – Final Presentation**

**May 2018**

**Junior Ovince**

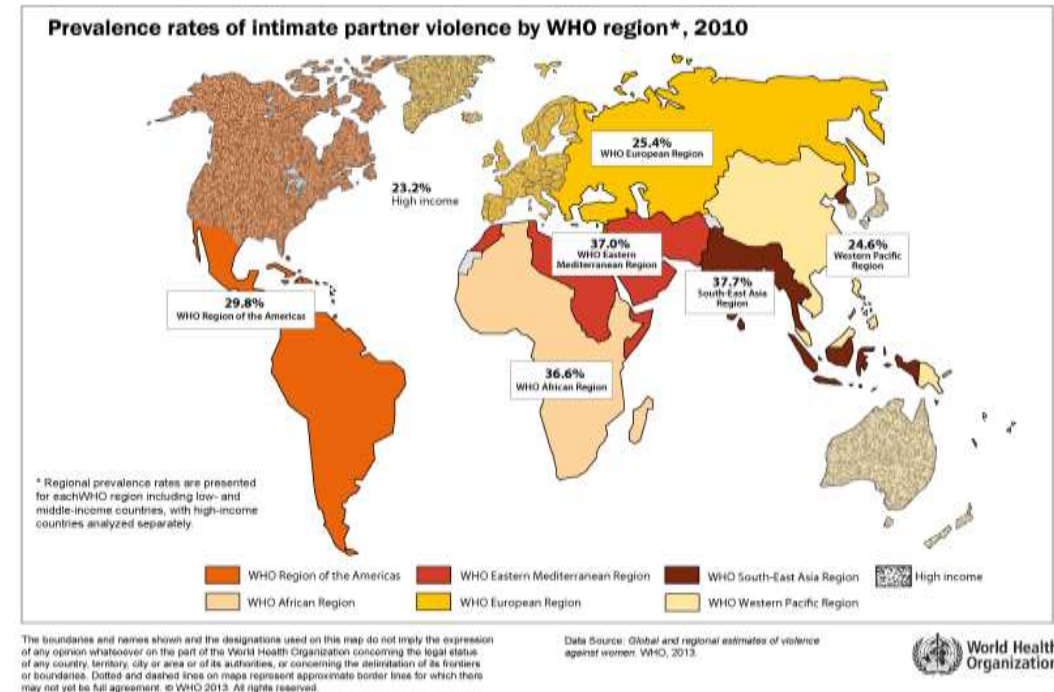
# Presentation outline

1. Problem Statement
2. Motivation and Objective
3. Dataset Description
4. Methodology
5. Results
6. Conclusion

# 1. Problem Statement

VAWG is a major public health problem and a crucial women's human right violation. According to a recent WHO report in 2013:

- A large proportion of the women's world, 35 %, have experienced either physical and/or sexual IPV or NPSV.
- Most of this violence are perpetrated by an intimate partner. Worldwide, one out of three ever-partnered women have suffered physical and/or sexual IPV. In some regions, this rate can be as high as 38%.
- Globally, in 38% 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.



# 1. Problem Statement (2)

The health consequences of physical and/or sexual IPV are alarming:

- 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

## Mental Health

**TWICE**  
as likely to experience depression




**ALMOST TWICE**  
as likely to have alcohol use disorders




## Sexual and Reproductive Health

**16%**  
more likely to have a low birth-weight baby



**1.5 TIMES**  
more likely to acquire HIV and 1.5 times more likely to contract syphilis infection, chlamydia or gonorrhoea



## Death and Injury

**42%**  
of women who have experienced physical or sexual violence at the hands of a partner have experienced injuries as a result



**38%**  
of all murders of women globally were reported as being committed by their intimate partners



# 2. Motivation and Objectives

## Motivation

A range of actions and efforts have been carried out both to prevent IPV from happening in the first place and to provide necessary services for survivors.

There is an opportunity to use machine learning techniques to inform **preventive** (better features) and **responsive** (better target) actions

## Objectives

- Identify the best classifiers to use for a pattern recognition of potential survivors of violence
- Improve the overall recall and f1-score of such classifiers to reduce false negative.

# 3. Dataset Description

## Data Source

- In the past 2 years, a representative national population survey on women's health and life experience using the WHO questionnaire and methodology was conducted in 5 countries in the Caribbean region: Jamaica, Trinidad and Tobago, Suriname, Guyana, and Grenada.
- These studies aimed mainly to determine the national prevalence rate of IPV and non partner violence (NPV) among women aged between 15 and 64 years old.





# 3. Dataset Description (2)

## 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 any partner (current or former) 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.

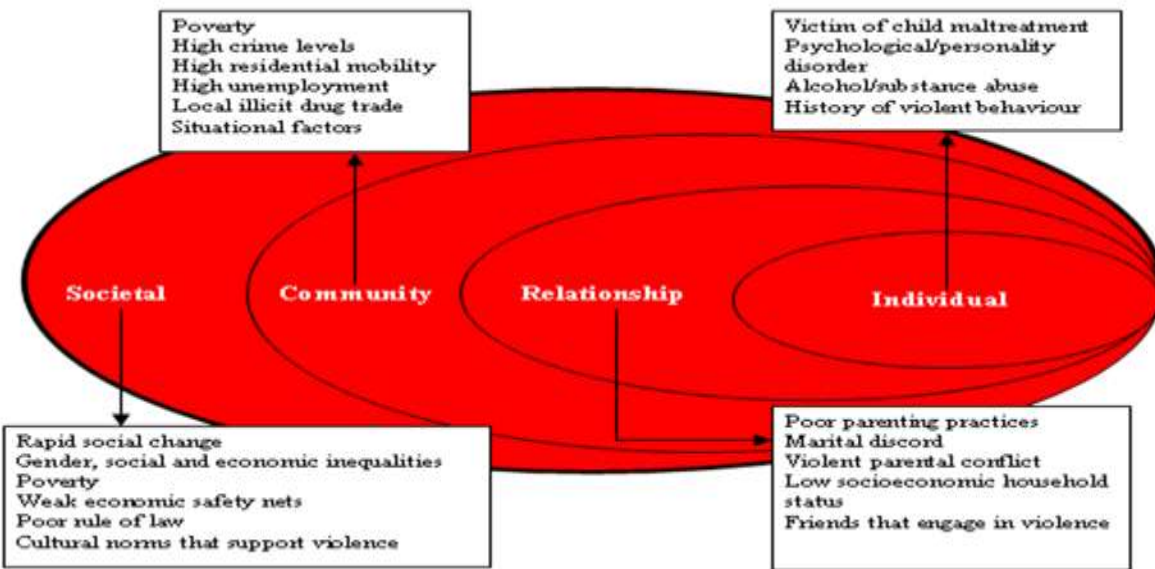
Table 1. Summary of the Data from the Caribbean National Surveys

	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
<b>Total</b>	<b>6250</b>	<b>5795</b>	<b>92.7%</b>	<b>32.3</b>

# 3. Dataset Description (3)

## Features

Based on the ecological framework

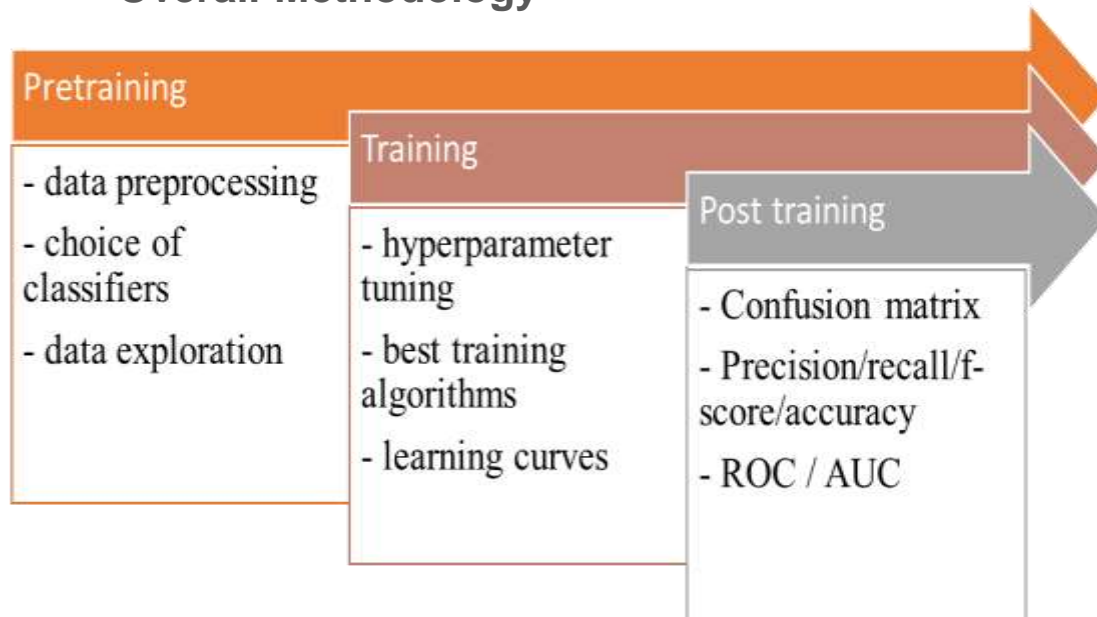


Category	Examples	Levels
Target	lifetime physical and/or sexual IPV (1 variable)	binary
Women's socio-demographic characteristics	age, religion, education, ethnicity, economic activities (7 variables)	individual
Women's social support	family support (1 variable)	relationship
Women's parenting and child 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



# 4. Methodology

## Overall Methodology



## A. Pretraining

### Data Preprocessing

- Cleaning and recoding
- Checking for missing values
- One-hot encoding of nominal/ordinal variables
- Splitting the data into train, validation, and test sets
- Normalization
- Features selection (using chi-square contingency and logistic regression)

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

# 4. Methodology / pretraining

## Choice of Classifiers

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

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)

# 4. Methodology / pretraining (2)

## Data Exploration

- 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.
- Finally, it is important to notice that country is not a significant feature.

Variables	witout other IPV	with other IPV
Intercept	0.033463	0.020439
emotvio (emotional IPV)	-.1	9.241461
econviol (economic IPV)	-.1	1.466971
Q515R (partner has other relationship)	1.764911	1.364926
Q513R (partner involves in fight)	1.704207	1.468292
mcv1006 (woman witnessed IPV at home)	1.562833	1.468292
mcv1006a (Women beaten in childhood)	1.352967	1.156386
mcv1006b (Women emotionally abused in childhood)	1.454264	1.184594
mcv1008 (partner witnessed IPV at home)	1.535107	1.448459
men_alcohol_all (man drinks alcohol at least once/week)	1.264529	1.148435
mcv1009 (Partner was hit as a child)	1.348644	1.22728
Fam_support (family support)	0.632737	0.684819
rQ701b (she shares things that happened to her during the day)	0.690044	0.67848
sQ601d (It is natural that men should be the head of the family_	1.371493	1.360973
CONTROLNUM_1 (one act of controlling behavior) none is reference	1.363834	1.188153
CONTROLNUM_2 (2 acts)	2.160414	1.469761
CONTROLNUM_3 (3 acts)	3.453886	2.165822
Q702R_2 (quarelling sometimes) never is reference	1.431038	1.202978
Q702R_3 (quarelling frequently)	3.391255	2.062255
earlymarriage_1 (married at 18 or younger)	1.59281	1.472262
edpart_1 (at most primary level) higher/tech/vocational is reference	1.399059	1.396683
sumdiffage_1 (woman is older) Partner at least 9 years older reference	1.387356	1.54311
country_1 (Guyana) Jamaica is reference	1.388605	1.014098
EP3_2 (living with man, not married) currently married is reference	1.557217	1.355811
EP3_3 (regular partner, living apart)	1.757513	1.466678
FCMAR_0 (non consensual relationship -no) not married is reference	1.32194	1.240358
FCMAR_1 (non consensual relationship -yes)	1.30787	1.266174
SourceIncome_2 (Income from own work)	1.351209	1.334291
SourceIncome_5 (Support from relatives/friends)	1.320486	1.451068
ageyr10_2 (25-34) 15-24 is reference	1.651031	1.617853
ageyr10_3 (35-44)	1.949943	1.707278
ageyr10_4 (45-54)	1.77039	1.48127
ageyr10_5 (55-64)	1.918033	1.34851

# 4. Methodology / 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.
- The SVM classifier in scikit-learn did not achieve convergence.

Summary of the Hyperparameters Tuned during the Training

Classifiers/Models	Hyperparameter Tuning	Scoring Sea
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

# 4. Methodology / post-training

## The traditional classification metrics

- Confusion Matrix: Globally informs on the classification accuracy and errors.
- Precision: is the ratio of true positive and the sum of true positive and false positive.
- Recall: is the ratio of true positive and the sum of true positive and false negative.
- Fscore: is the harmonic average of precision and recall.
- Accuracy: is the ratio of the number of classes correctly classified and the total of samples.
- 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.

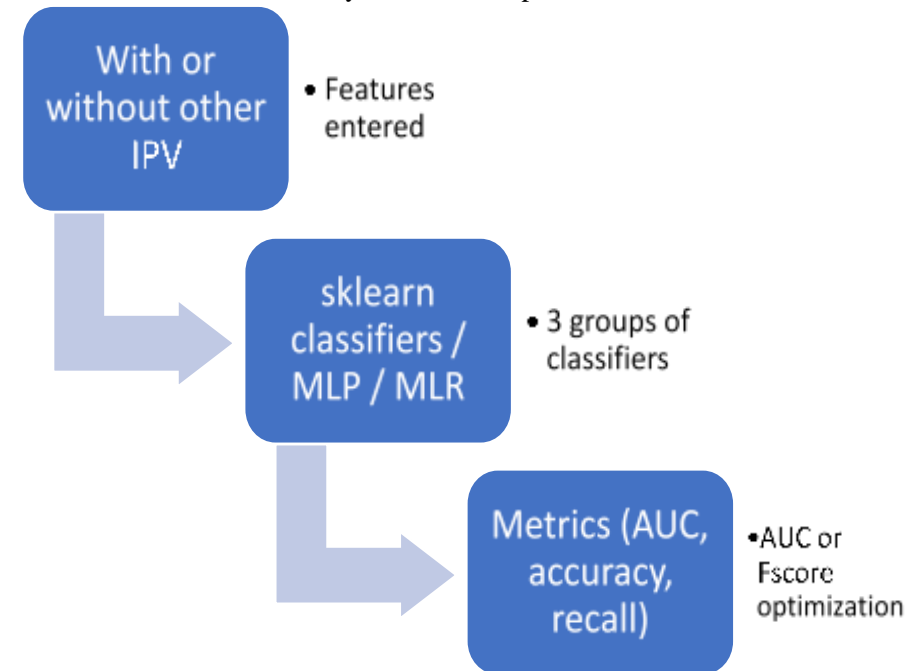
		Predicted class	
		$P$	$N$
Actual Class	$P$	True Positives (TP)	False Negatives (FN)
	$N$	False Positives (FP)	True Negatives (TN)

# 5. Results

## The traditional classification metrics

- 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 not presented.

Summary of the results presentation

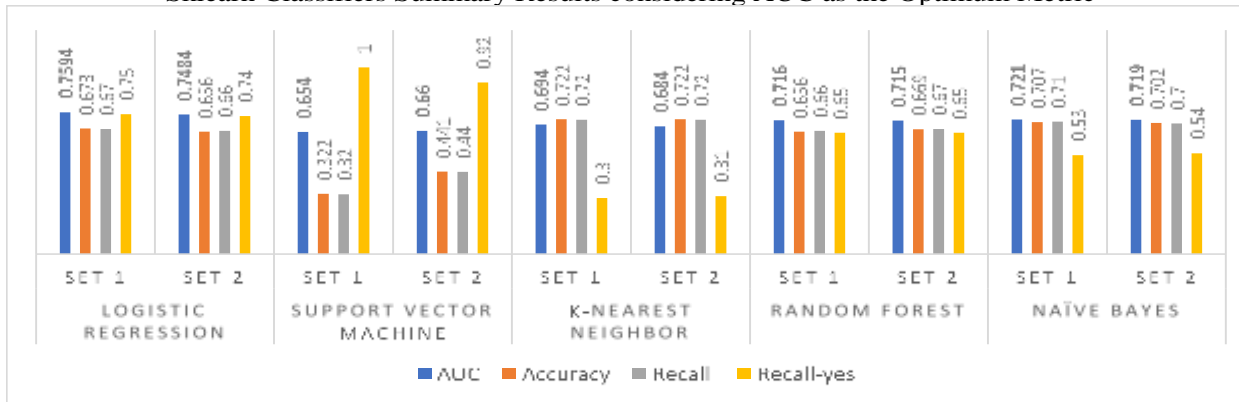




# 5. Results / without other IPV

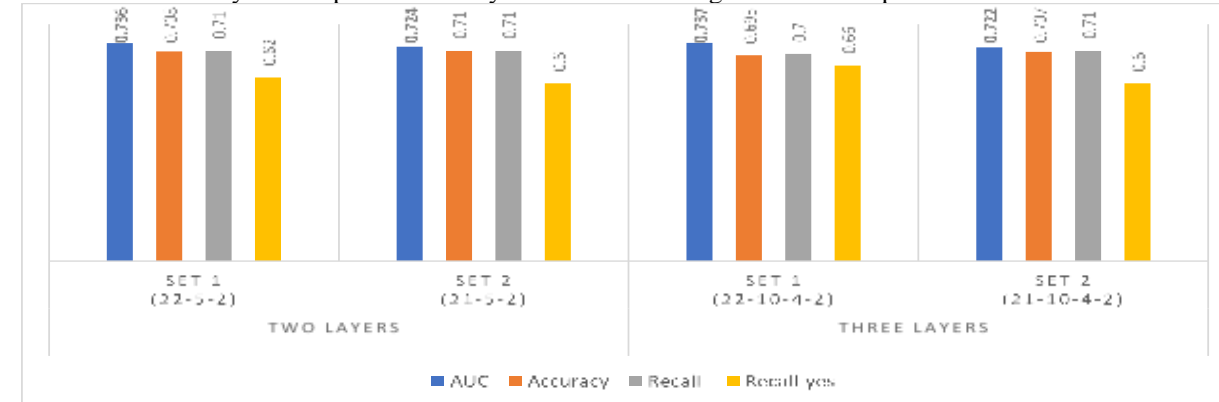
## The sklearn classifiers

Sklearn Classifiers Summary Results considering AUC as the Optimum Metric

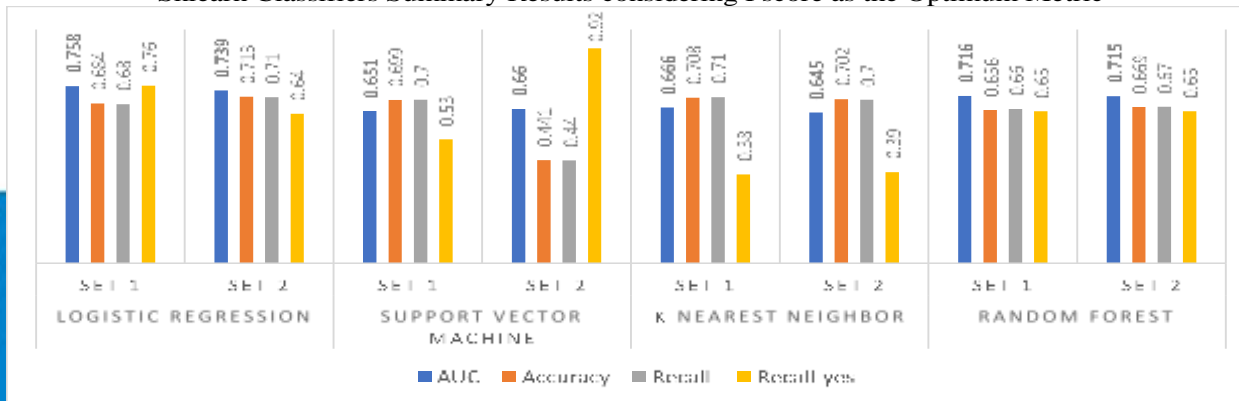


## The mlp network

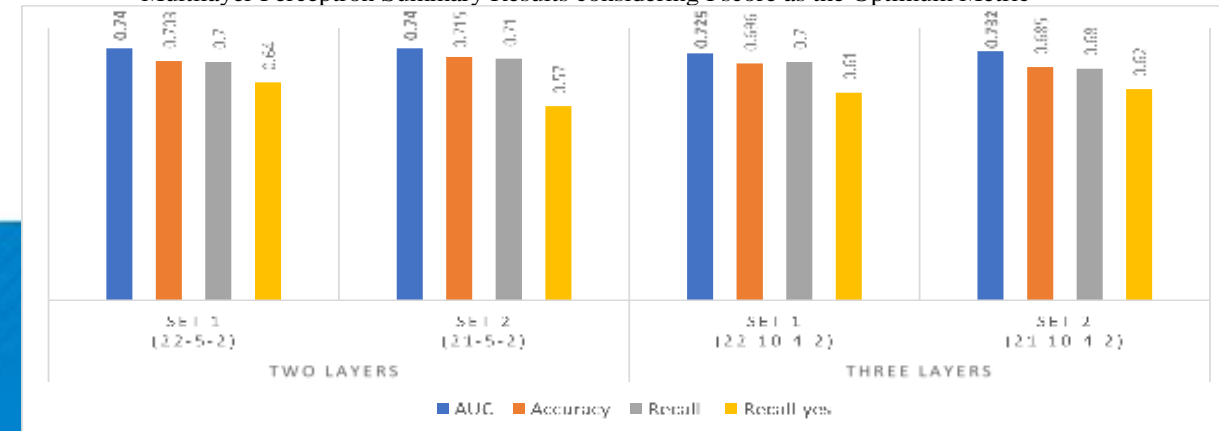
Multilayer Perceptron Summary Results considering AUC as the Optimum Metric



Sklearn Classifiers Summary Results considering Fscore as the Optimum Metric



Multilayer Perceptron Summary Results considering Fscore as the Optimum Metric



# 5. Results / without other IPV (2)

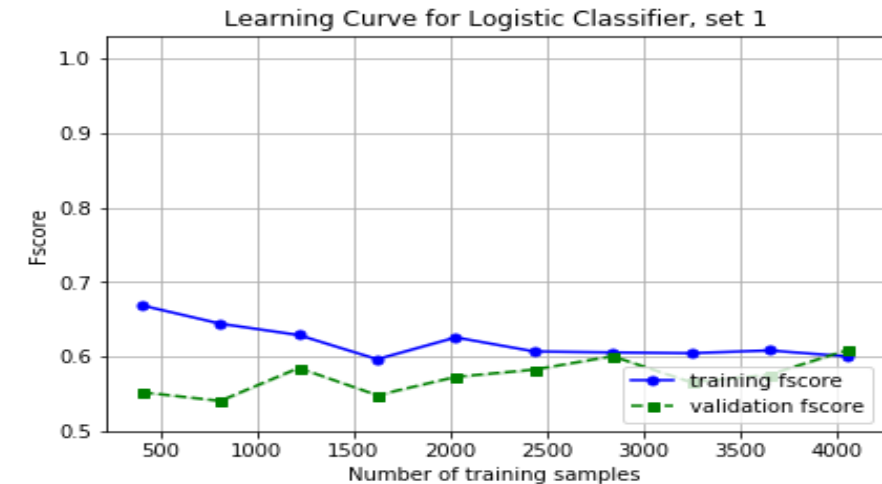
## The multilevel regression

- 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, 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).

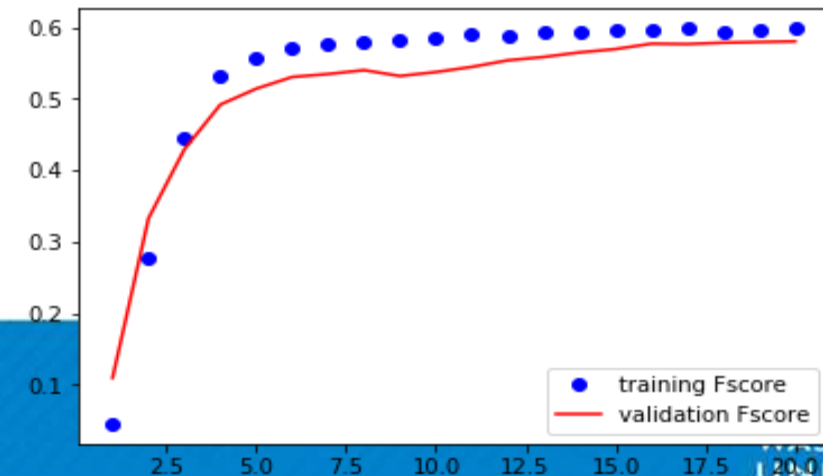
Accuracy : 0.7526
95% CI : (0.7225, 0.781)
No Information Rate : 0.6778
P-Value [Acc > NIR] : 8.428e-07
Kappa : 0.3691
McNemar's Test P-Value : 1.764e-13
Sensitivity : 0.4214
Specificity : 0.9100
Pos Pred Value : 0.6901
Neg Pred Value : 0.7679
Precision : 0.6901
Recall : 0.4214
F1 : 0.5233
Prevalence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

# 5. Results / without other IPV (3)

- 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 or the training for the 2-layer network achieves 0.6 as the highest score.
- 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.



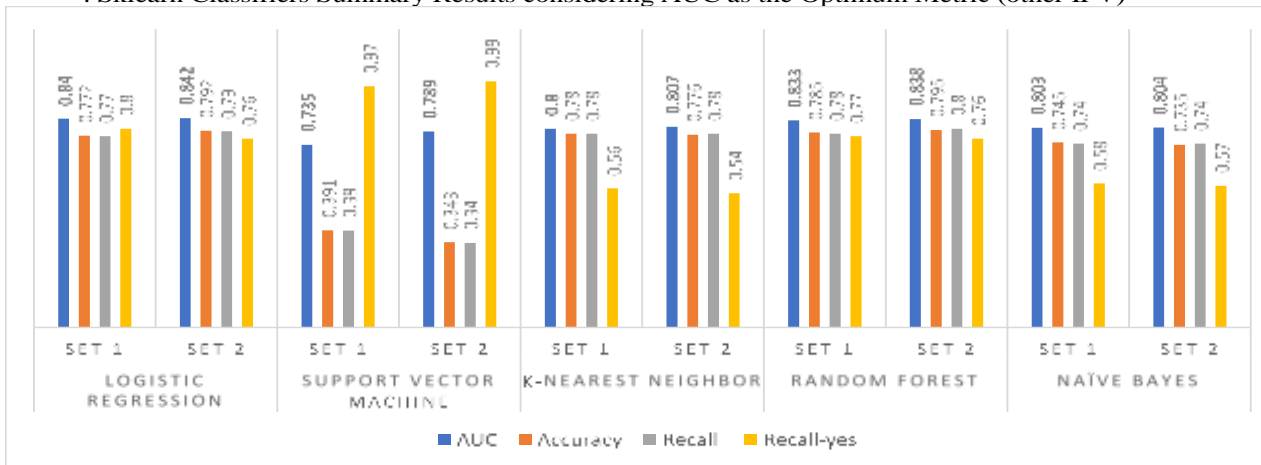
. Training Performance of the 2-layer Network (fscore and set 1)



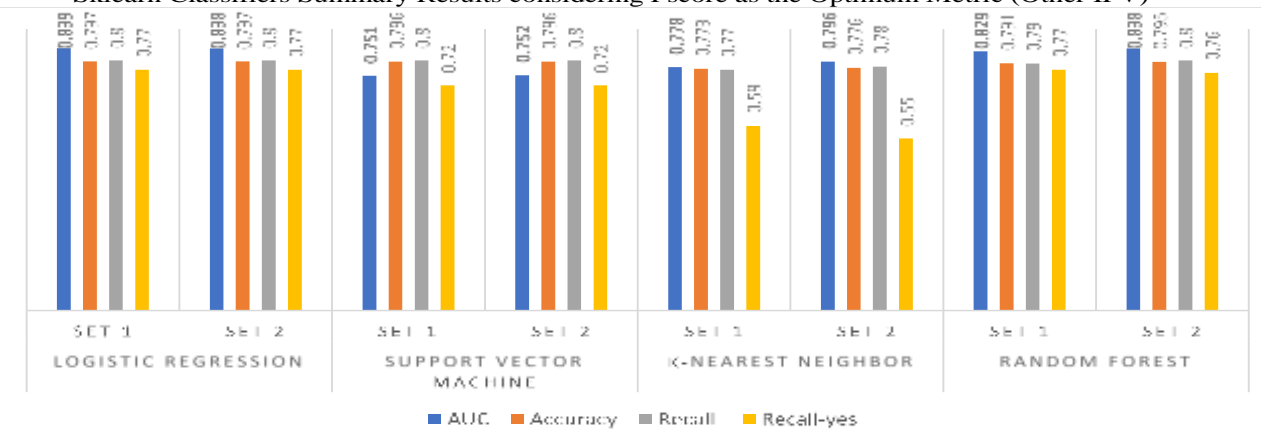
# 5. Results / with other IPV

## The sklearn classifiers

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

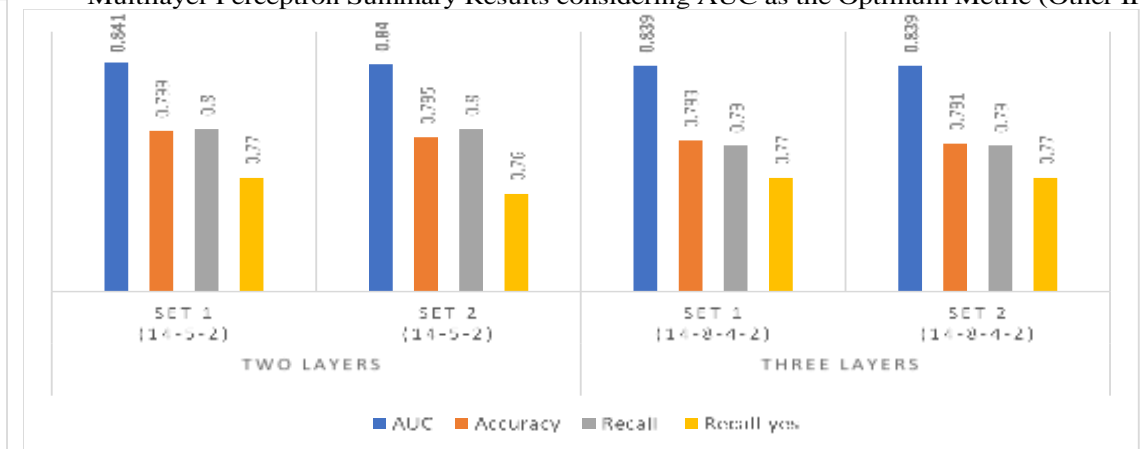


Sklearn Classifiers Summary Results considering Fscore as the Optimum Metric (Other IPV)

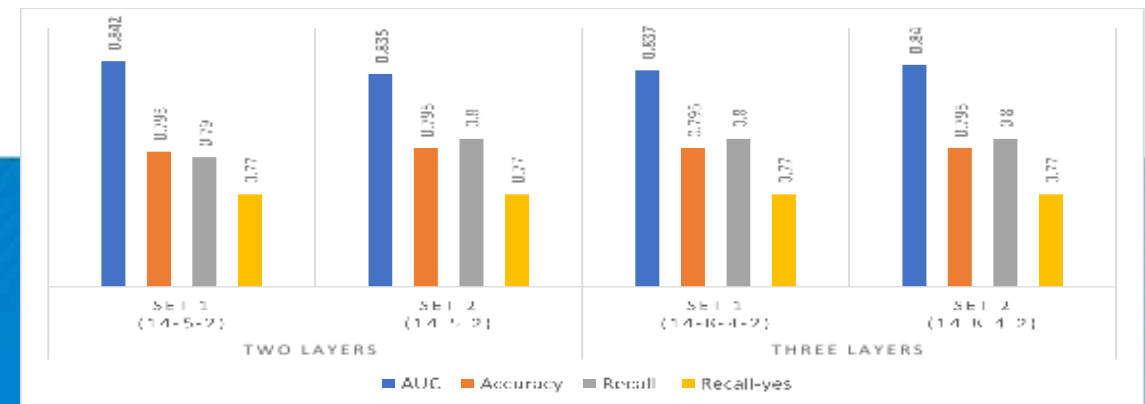


## The mlp network

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



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



# 5. Results / with other IPV (2)

## The multilevel regression

- 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.

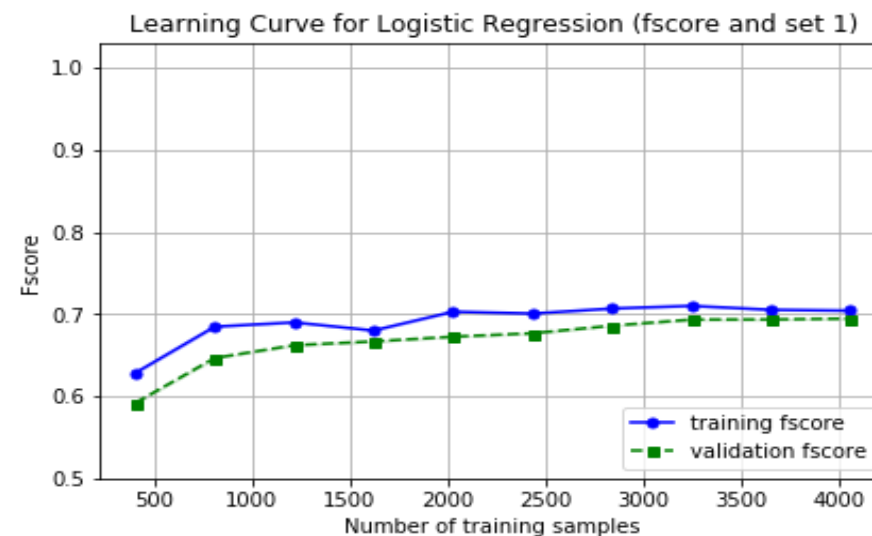
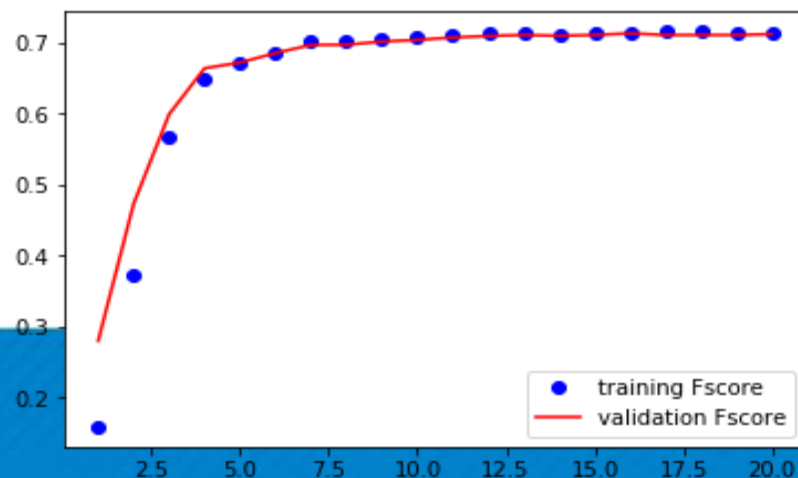
Multilevel Regression Summary Results (with other IPV)

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

# 5. Results / with other IPV (3)

- 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.

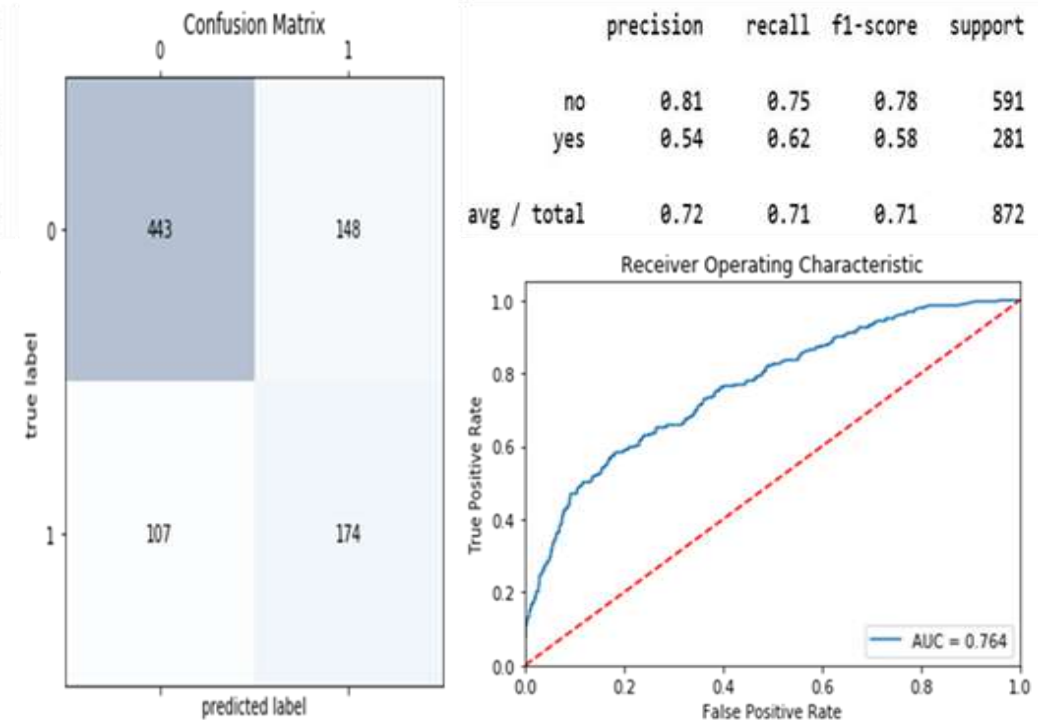
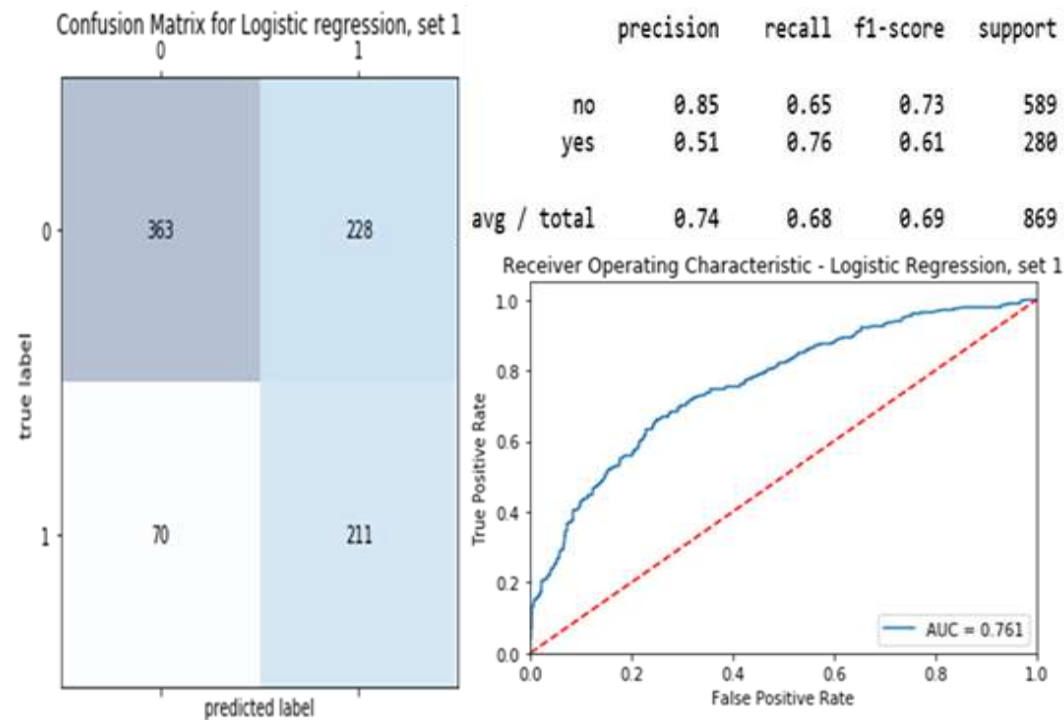
. Training Performance of the 2-layer Network (other, AUC, and set 1)





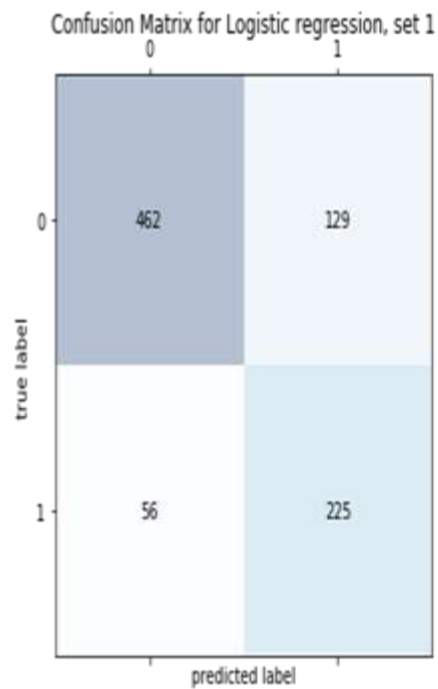
# 5. Results / Model Evaluation

Without other IPV

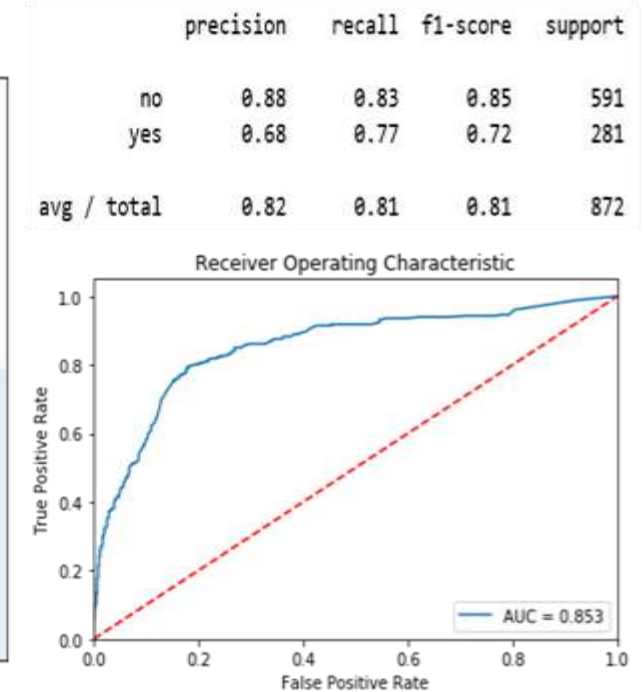
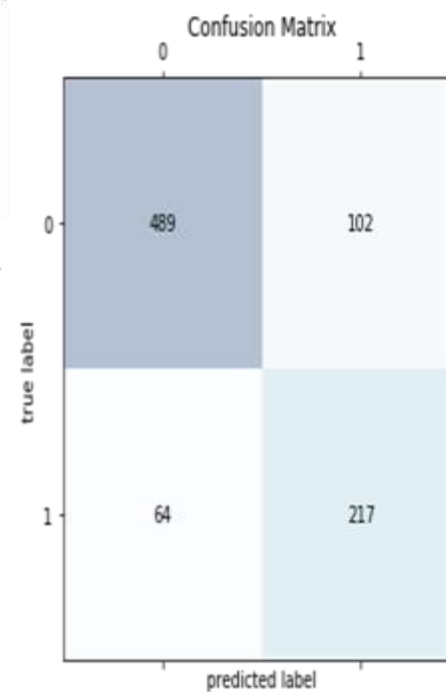
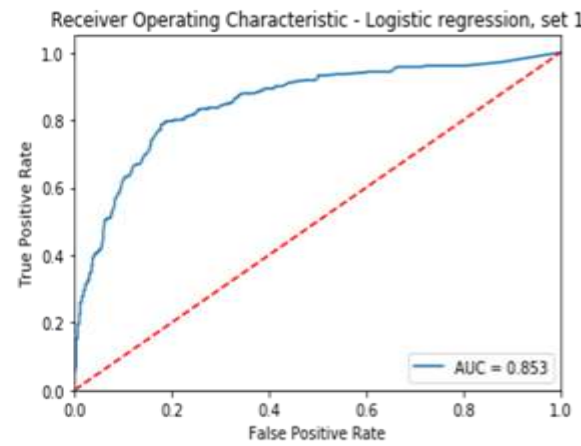


# 5. Results / Model Evaluation (2)

With other IPV



	precision	recall	f1-score	support
no	0.89	0.78	0.83	591
yes	0.64	0.80	0.71	281
avg / total	0.81	0.79	0.79	872



# Conclusion

## Summary

- 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 two 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.

## Future research considerations

- Feature scaling
- Adding other features
- Hierarchical modeling
- Inclusion of interaction

## Limitations

- Self-reporting of IPV

---

**THE GEORGE  
WASHINGTON  
UNIVERSITY**

---

WASHINGTON, DC

