Predicting Gender Based Violence in Mexico: A Machine Learning Approach

Ana Sofia Munoz (anasofiamvaladez), Gabriela Palacios (gabrielapg), Antonia Sanhueza (asanhueza1) and Hugo Salas (hsalasr)

Executive summary:

A recent estimate suggests that 30% of all women worldwide have suffered some form of gender based violence. However, most of these cases remain hidden from the public eye, as they are often not reported to the relevant authorities. This makes it difficult for policy makers to provide support services to women in danger, as cases are hard to detect. With the hope of overcoming this constraint, we used individual-level survey data to predict violent episodes suffered by women in Mexico. We trained three supervised machine learning algorithms for this purpose: a logistic regression, a decision tree and a random forest. Our results showed that our classification algorithms lacked the power to effectively identify women who have suffered from GBV with their current partner. However, among all our models, Logistic Regression outperformed the rest. Further research needs to be performed in order to obtain a reliable estimate of GBV at the individual level.

Repository :

https://drive.google.com/drive/folders/1_JB-21oEyYrg89R-MrHqg3U2nj8uHqJw?usp=sharing

1. Background and Overview

The World Health Organization (WHO) estimates that 736 million women (about 30% of all women) have been subjected to intimate partner violence, non-partner sexual violence, or both at least once in their life.[1] In Mexico, the problem might be particularly disturbing: the National Institute of Statistics (INEGI, for its acronym in spanish) reports that 66% of women older than 15 have suffered some sort of physical or sexual violence, 17% have suffered from a physical assault and 11% has been victim of sexual violence in an academic environment.[2]

Despite its relevance, Gender Based Violence (GBV) is very hard to tackle: compared to other forms of violence, it is less likely to be reported and, therefore, it is hard to detect. Underreporting happens for several reasons: fear, since victims might continue to see her victimizers regularly; social discredit, as accusations might be associated with social stigma; lack of trust in the judicial system; and/or the high costs that come by initiating a judicial process, among others (April, 1999; UNICEF, 2000; Watts and Zimmerman, 2002). According

[1] World Health Organization, on behalf of the United Nations Inter-Agency Working Group on Violence Against Women Estimation and Data (2021). Violence against women prevalence estimates, 2018. Global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence against women.

[2] Instituto Nacional de Estadística y Geografía (México). Panorama nacional sobre la situación de la violencia contra las mujeres / Instituto Nacional de Estadística y Geografía.-- México : INEGI, c2020.

to INEGI, in 2016 only 9.4% of all the violent events suffered by women were reported to the authorities. Therefore, focusing on supporting women who have reported a case of violence would leave a great majority of the victims without help.

An additional layer of complexity arises when we consider the collection of information regarding GBV: the deployment of a quality survey requires high-level training, very strict interview protocols and large monetary investments. This is probably one of the main reasons why, as UNDP reports, such surveys are not sufficiently implemented in Latin America.[3] Nevertheless, INEGI collects one of the only GBV-focused surveys in the region that adheres to the highest measurement standards: the National Survey about the Dynamic of Relations in the Household (ENDIREH for its acronym in spanish). This survey characterizes violence at the individual level (women) along with information about the household sociodemographic characteristics. One of the main advantages of ENDIREH is that it also contains information about perceptions on gender roles, economic independence, and trust, which have been found to be highly correlated with GVB (WHO, 2009; Begum et al, 2017). [4] [5]

Using microdata from ENDIREH, we trained a set of Machine Learning models that aimed to predict if a woman is a victim of GBV. In particular, we hoped to produce a tool that could help governments and local NGOs identify women at risk, in order to provide them with resources and accompaniment to attack the problem.

With this tool, survey designers, public policy makers and NGOs that are trying to tackle GBV will be able to produce better GBV-focused policies and protect women more effectively. By reducing the number of questions/costs needed to assess domestic violence in a questionnaire, we hope that policymakers will be more likely to prevent and act upon it. Furthermore, by providing the main variables that predict GBV we expect that laws and programs that protect women become more effective and better targeted.

2. Data

For this project, we utilize data from the National Survey about the Dynamic of Relations in the Household (ENDIREH for its acronym in spanish). This is a national and subnational representative survey that is collected every 5 years in Mexico. It is only asked to women aged over 15 years of age and is heavily focused on questions regarding past violent events and gender roles. In order to ask such sensitive questions, the survey follows a set of strict procedures:

- Only women serve as enumerators;
- All enumerators are well trained as social workers;

[3] From Commitment to Action: Policies to End Violence Against Women in Latin America and the Caribbean. Sebastián Essayag. Panamá: UNDP and UN Women 2017.

[4] Begum, S., Donta, B., Nair, S., & Prakasam, C. P. (2015). Socio-demographic factors associated with domestic violence in urban slums, Mumbai, Maharashtra, India. *The Indian journal of medical research*, *141*(6), 783.

[5] World Health Organization. (2009). Promoting gender equality to prevent violence against women.

- Sensitive questions are registered in an electronic device directly by the women surveyed, which guarantees full confidentiality;
- Women are required to be surveyed when alone (even if that requires accompanying her to the supermarket or to any daily chores)

Although most questions available are at the individual (woman) level, it also collects relevant information about demographics of household members.

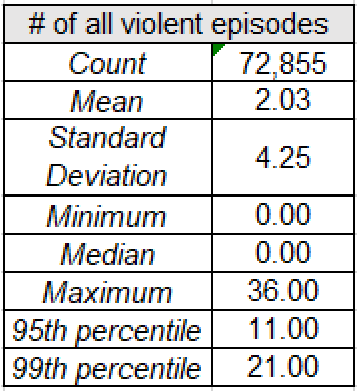
ENDIREH asks four different questionnaires depending on the type of respondent: (1) woman with a partner, (2) widow, (3) separated or (4) single. Our analysis will focus only on questions directed to women in a relationship (either married or in a free union), since they represent the group with the highest share.

For convenience, our unit of analysis is the woman and we have aggregated all household information to that level. This means that each observation/row contains information about one woman and its household characteristics, such as household size, share of women with respect to men, age of household members, household assets, education attainment and relevant information of the woman's partner. Our final dataset contains 72,855 observations/women/rows.

Our main goal for this project is to predict Gender Based Violence (GBV). Therefore, our first task was to create a measure of GBV using our data. For this, we chose to focus on a module that asked women about violent events that happened during their current relationship.

Under this module, there are 36 different questions of this nature. As an introductory step, we aggregated all these measures into one general variable that measures the number of different violent events that each woman reported (number of questions for which the women's answer was "Yes"). Figure 1 shows the histogram of such a measure. As we can see, nearly 60% of our respondents did not suffer from any violence during their current relationship.

Table 1: Summary Statistics



# of all violent episodes	
Count	72,855
Mean	2.03
Standard Deviation	4.25
Minimum	0.00
Median	0.00
Maximum	36.00
95th percentile	11.00
99th percentile	21.00

The table above summarizes this distribution further. Moreover, as the nature of our purpose is classification, we decided to collapse our variables so that we have women that have not suffered from any type of violence (= 0) and those that have suffered any type of violence (>0).

This left us with a categorical variable equal to 1 for women who have suffered from any type of violence (39.4% of the sample) and 0 for those that have not (60.6% of the sample).

Finally, Figure 2 shows the geographic distribution of our newly created target variable with respect to Mexico's subnational divisions (states). As it can be seen, prevalence rates are highly concentrated between 30% and 50% with the highest prevalence being concentrated in the black states near Mexico's center. Coincidentally, this is also Mexico's most populous metropolitan area: Mexico City and the State of Mexico.

3. Machine Learning and Details of Solution

3.1 Data Preprocessing

The following processes were implemented in order to prepare our data for our machine learning models:

I. **Construction of a target variable:** Based on Section XIII of questionnaire A, which asks about a woman's violent episodes with her partner, a dummy variable called 'suffers_violence' was created. '1' indicates that she suffered at least one violent event and '0' otherwise.

II. **Construction of feature variables:** Based on the general questionnaire and questionnaire A available for ENDIREH, we put together a set of features that we thought would be relevant for the prediction task at hand. In terms of household level characteristics: we gathered sociodemographic (age, gender, education levels, literacy, of all members in the household), labor, income, dwelling and asset information. Regarding the woman's characteristics, we gathered data about her habits, her support circle, opinions about gender roles and actual distribution of chores around the house.

III. **Missing value imputation**: However, for most of our features, missing values were more or less around 0.03% of the dataset. We used the following rules to deal with missing values:

   A. Categorical variables: the mode of the sample was imputed.
   B. Continuous variables: the median of the sample was imputed.
   C. Cases where missing values represented more than 5% of the data: this happened for the women for which we couldn't identify her partner's characteristics. For those cases, since they were all categorical, we treated the missing values as if they were an additional category.

3.2 Feature engineering

Before including all of our variables into our model, we performed the following tasks:

I. **Expanded categorical variables into dummies**: as it is common from household surveys, most of our available features were categorical. In order for these categorical variables to be used into our models, we transformed them into one dummy variable for *k* - 1 (*k* being the number of categories). This step expanded significantly the dimensionality of the data set; but since *sci-kit learn* cannot handle categorical data for decision tree models (it treats them as continuous, which causes a loss of information), we thought it was the only alternative we had.

II. **Polynomials**: A function was created to calculate the interaction between features, the second degree polynomial of every feature, and their correlation with the target variable. The latter, to detect new variables which were worth adding. However, we found little correlation among interactions and target variables, so none of them were included.

III. **Removing highly correlated and constant variables*:*** we removed features that had multicollinearity problems between each other. For instance, partner's age and the woman's age are linear combinations of the differences between both, so one of these was dropped. Constant variables such as gender were removed, since the whole sampel is female.

3.3 Machine Learning Problem and Model Selection

We restricted the problem to that of identifying the best features that predict whether a woman suffers any type of GBV or not, making our machine learning model a classification problem. It is worth noting that ENDIREH holds a wide variety of violence questions; it even has records of different types of violence (psychological, economic, sexual and physical) and in different environments (the community, schools, within the household, among others). Instead of digging deep into each of these types, we decided to look at aggregated violence data (of all types) within the household. We believe that, from a policy perspective, all types of violences should be treated with importance. Moreover, this is a first glance at the problem and we hope to be able to expand this research in the future.

In terms of models, we chose to train three machine learning models: (Single) Decision Tree, Logistic Regression and a Random Forest. We chose these models because they are widely used in classification algorithms that have proven to be successful in several prediction problems. Additionally, we ran an exploratory supervised analysis using a K-modes classifier to see if the clusters that resulted from this exercise showed differential violence rates.

We started with a simple tree because it is computationally less expensive than other models and is very intuitive, thus easy to explain. After that, we used the same hyperparameters for the estimation of the Random Forest. One relevant advantage of this technique is the ability to determine variable importance based on each variable's contribution to an increase in purity and a decrease in error. Furthermore, we attempted to use clustering as an unsupervised approach, but because of the difficulties that imply the use of clustering in a high dimensional data set, we applied principal component analysis before its use. A caveat to this approach is that it is not recommended under categorical values since PCA is designed for continuous variables as it tries to minimize variance and results are less meaningful for categorical variables.

All of our models were trained using 80% of the data as training set and 20% as testing set. Except in the clustering approach, where we worked with a random sample of 50 per cent of the entire feature set; our computers were not capable of running the algorithm due to lack of memory.

We tuned the hyperparameters of the Tree and Random Forest by using a 10-fold Cross Validation approach and added the maximum number of estimators to the hyperparameters. To choose among all the hyperparameters, we used the recall scoring metric. We chose this metric since we consider that it is more important to detect as many victims of GBV as we can, even if that comes at a cost (false positives).

4. Evaluation and Results:

4.1 Logistic Regression

We took off our classification task with a logistic regression. This is a fairly simple and easy to interpret non-linear model that is based on the logistic (sigmoid) function. For this particular model, all of our continuous variables were transformed into quintiles and later coded into different dummy variables, in order to have all features on the same scale and avoid standardizing dummy variables.

To penalize the regression and avoid overfitting, an elastic net regularization was implemented using the 10-fold cross-validation. The most accurate model ended up having an alpha parameter of 0.0001 and an epsilon parameter of 0.1. We also used a balanced approach to weight observations such that the share of classes between the target variable were equal.

4.2 Tree

Next, two classification tree models were produced (see Anex X). One using default class weights, and balanced class weights. Balanced class weights adjust weights inversely proportional to class frequencies. The best performing tree in both models used the Gini criterion, had a minimum sample to split of 100, and a maximum depth of 30. This model yields a 37% recall score in the training sample, and 37% in the testing sample.

The balanced tree outperformed the regular tree model. Thus, we only present results for the balanced case.

4.3 Random Forest

Just like in 4.2 we ran two Random Forest Models, one with regular class weights, and one with balanced class weights. The latter model also outperformed the regular Random Forest. Thus we only present results for the balanced case.

The best performing model in the training sample used 100 trees, used the Entropy criterion, was split at a minimum of 1,000 samples, and had a maximum depth of 20.

The feature importance of this model is shown in Figure 3. According to this model, the most relevant questions to determine if a woman is suffering GBV are related to i) the people in which they support in case of problems, ii) the monthly income of the woman, and iii) gender role perceptions.

4.4 Interpretability of Principal Component Analysis and Clustering

The assumption is that the ENDIREH data set includes highly correlated features and that there exists a subspace with smaller dimensions that can provide nearly the same information as the original data.

After detecting the components which captured the greatest amount of variance, a correlation matrix was calculated between the principal components and the features to distinguish the main variables of which the main component was composed. Using these variables we estimated a K-Modes Clustering with the features: residence in mother’s house, 'Relationship of the woman with the head of the family', ‘Bachelor degree’ and 'Age difference'. The standard K-means algorithm isn't directly applicable to categorical data because the sample space for categorical data is discrete and an Euclidean distance function on such a space isn't intuitively meaningful. Instead we introduce the variation of k-means more suitable for mixed data (categorical and numeric), which uses a simple matching dissimilarity measure to deal with categorical objects and replaces the means of clusters with modes, and uses a frequency-based method to update modes in the clustering process to minimize the clustering cost function.

As opposed to classification, it is difficult to assess the quality of results from clustering. We decided to measure the goodness of the split by calculating the incidence of each cluster group in the original labels. However, as shown in the table below, results suggest that the variables selected by de PCA analysis are not good predictors of violence.

Table 2: Incidence

	Doesn’t Suffer Violence	**Suffers Violence**
Cluster 1	0.602	0.597
Cluster 2	0.398	0.4026

4.5 Evaluation metrics

To evaluate the performance of the model we used recall as our main scoring metric because we want whoever is using the tool that all victims of GBV are detected. It is more important not to leave true positives outside than getting false positives. Below, there is a summary of the different models' scoring metrics. These models were chosen from the 10-fold Cross Validation as the ones that maximized recall.

Table 3: Summary scoring metrics

Model	Recall	Precision	Accuracy
Logistic	0.45	0.52	0.62
Tree	0.60	0.46	0.57
Random Forest	0.56	0.50	0.61

From the results, we observed that the Recall runs between 44 and 55%, Precision runs between 43 and 53%, and Accuracy runs between 57 and 62%. In other words, Recall shows only around 55% of those that are positive were classified as such; Precision shows around 50% of positive predictions are classified correctly; and Accuracy shows around 60% of predictions were classified correctly.

Maximizing Recall would lead us to choose Tree as the best model. However, when we look at the Precision-Recall Curves (Figure 4), we can tell that the Logistic model has a higher than or equal Precision for every level of Recall. Probably, the Tree model has a higher average due to the big spike at the beginning. Taking this into account, we conclude that Logistic Regression is the best model to maximize Recall.

Figure 5 shows the distribution of prevalence of violence based on this model's testing set. We can see that in fact, the boxplots overlap a lot, even medians are very close. Therefore, the models are not doing a good job at giving high probabilities to those that have suffered violence, and low to those that have not.

Additionally, we created another evaluation metric that is specific to our application. This metric could be useful for organizations who are interested in estimating the share of women that have suffered from any type of violent episodes; instead of having an accurate classification of women, we just want to know the proportion of women who are in a certain class. For instance, when Mexico’s biggest cash transfer program was first implemented (Progresa/Oportunidades/Prospera), the targeting criteria was at the locality level (only localities with certain levels of poverty were chosen to be part of the program) instead of the household level (although the delivery method was the household).

For this purpose, we performed a set of bootstrapped simulations. First, we randomly sampled 200 observations without replacement from our test set. Second, we calculated the real prevalence rate (% of women suffering from GBV) and our estimated prevalence rate (the average of all our predicted probabilities). Third, we repeated this process 5,000 times. The following table summarizes the results of this exercise. Consistent with our recall/precision curve, the logistic regression outperforms the other two models. Specifically, when estimating a prevalence rate in a sample of 200 women, we should expect our error to be between +1.80 and -12.21 percentage points of the true value 95% of the time when using a logistic regression.

Table 4: Model performance: Bootstrapped simulations

Model	Average error	Lower bound (95% CI)	Upper bound (95% CI)
Logistic	-5.15	-12.21	1.80
Tree	-9.99	-16.96	-3.22
Random Forest	-10.22	-16.96	-3.13

The feature importance of the logistic model in Figure 6 are similar to those of the Random Forest. In particular, we see that variables that explain the most variance come from the questions about division of household chores, support circle and economic independence.

5. Policy implications

This paper represents an initial step towards using machine-learning methodology to predict gender-based violence in Mexico. As Logistic Regression is our best model, we can use the importance of features to select the variables for designing a short questionnaire for the purposes mentioned above. The greatest weakness of this exercise is the low degree of recall for all of our models, in this sense this model cannot be used to detect a totality of cases.

The purpose of this exercise is to offer non-profits and/or government agencies a start point to design a screening test to detect violence against women from a small group of questions. This reduced number of important features can be used in places such as hospitals, which are the best places to make this kind of detections[6]. This tool could also help prevent cases of possible GBV by establishing the characteristics under which women tend to be victims of it.

6. Limitations, Caveats and Ethics

As we mentioned in the beginning it is hard to identify victims of GBV, and one of the barriers is the lack of reporting. According to the World Bank, fewer than 40% of women who experience violence seek help of any sort, and fewer than 10 percent seek help appealing to the police. Even though ENDIREH is an anonymous survey, and it is implemented in a safe environment with people trained to ask these questions, it is likely that some women still don't report violent episodes. This is a statistical bias we must be aware of when interpreting the results.

One of the biggest limitations of our model was that the high dimensionality due to the high number of categorical variables, made it very expensive computationally.

Our study aims to help identify women suffering GBV by establishing a subset of variables that are good at predicting women that are victims of GBV. In no case are we establishing a causal relationship between the answer to these questions and being victimized.

[6] K. Kero, A. Puuronen Usability of two brief questions as a screening tool for domestic violence and effect of #MeToo on prevalence of self-reported violence. 2020.

This is a very delicate issue that affects women's life in a very intimate way. Thus, the use of this tool must be done with extreme caution. Human subjects need informed consent. We must also protect the privacy of the information they provide, and of the possible treatment they might receive due to this.

References

Begum, S., Donta, B., Nair, S., & Prakasam, C. P. (2015). Socio-demographic factors associated with domestic violence in urban slums, Mumbai, Maharashtra, India. *The Indian journal of medical research*, *141*(6), 783.

Essayag, Sebastián. Panamá: UNDP and UN Women 2017. "Commitment to Action: Policies to End Violence Against Women in Latin America and the Caribbean."

Instituto Nacional de Estadística y Geografía (México). Panorama nacional sobre la situación de la violencia contra las mujeres / Instituto Nacional de Estadística y Geografía.-- México : INEGI, c2020.

K. Kero, A. Puuronen Usability of two brief questions as a screening tool for domestic violence and effect of #MeToo on prevalence of self-reported violence. 2020.

World Health Organization, on behalf of the United Nations Inter-Agency Working Group on Violence Against Women Estimation and Data (2021). Violence against women prevalence estimates, 2018. Global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence against women.

World Health Organization. (2009). Promoting gender equality to prevent violence against women.

Annex

Figure 1: Distribution of violent episodes

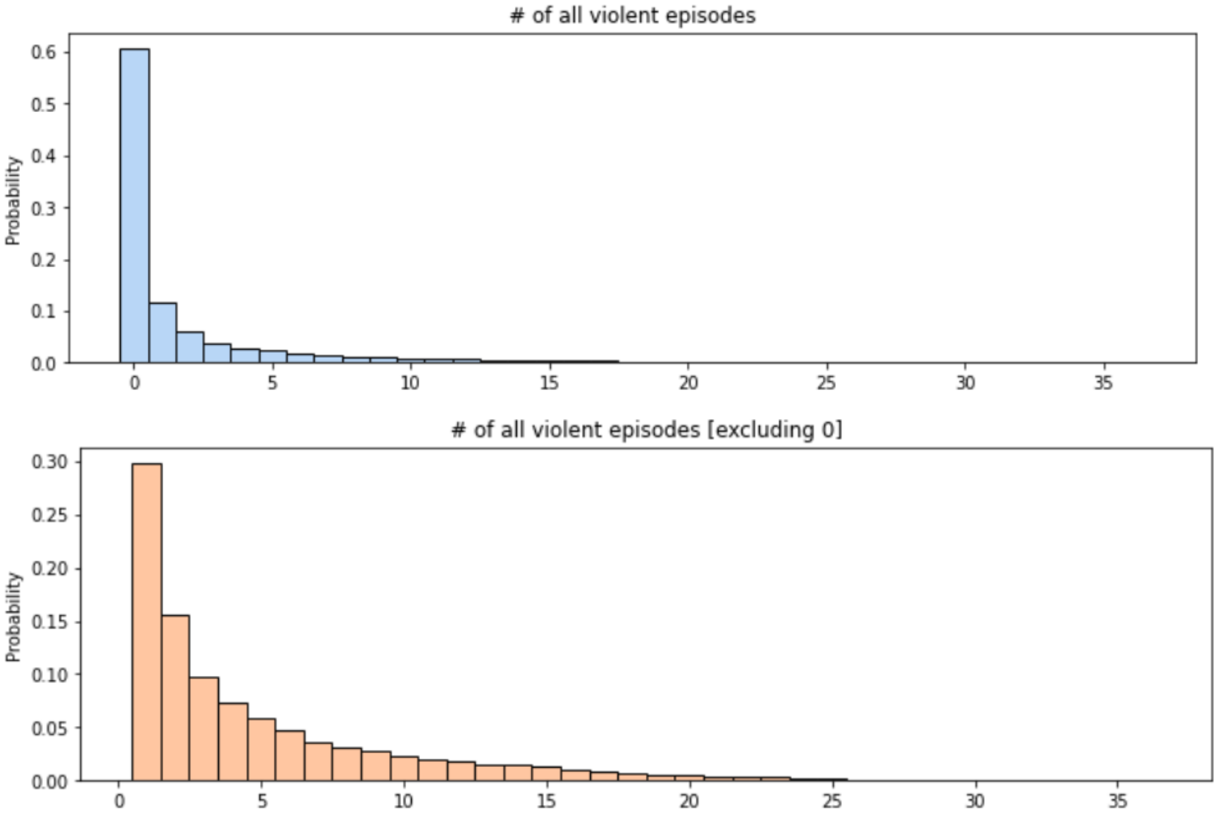


Figure 2: Geographic prevalence of GBV in Mexico

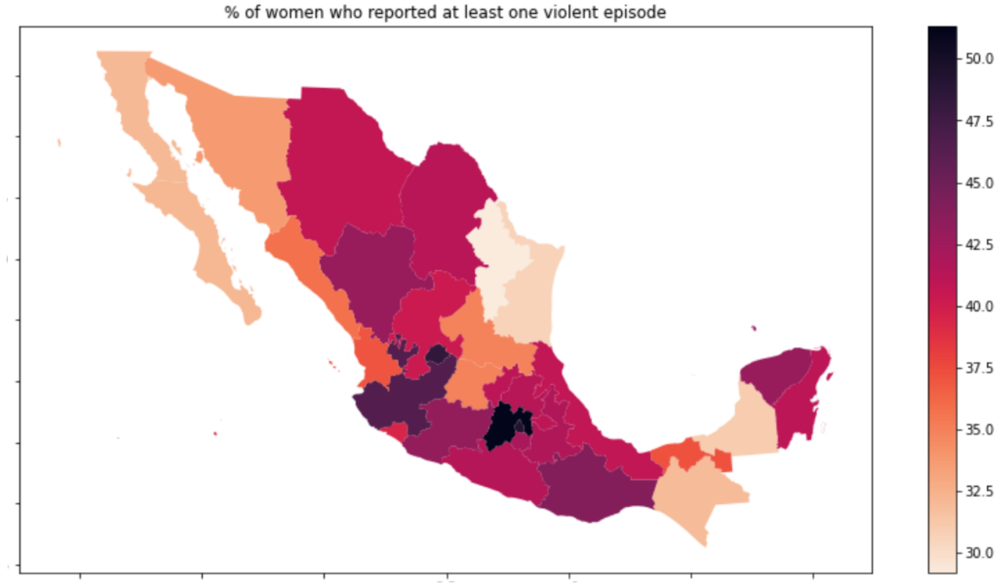


Figure 3: Random Forest Feature Importance

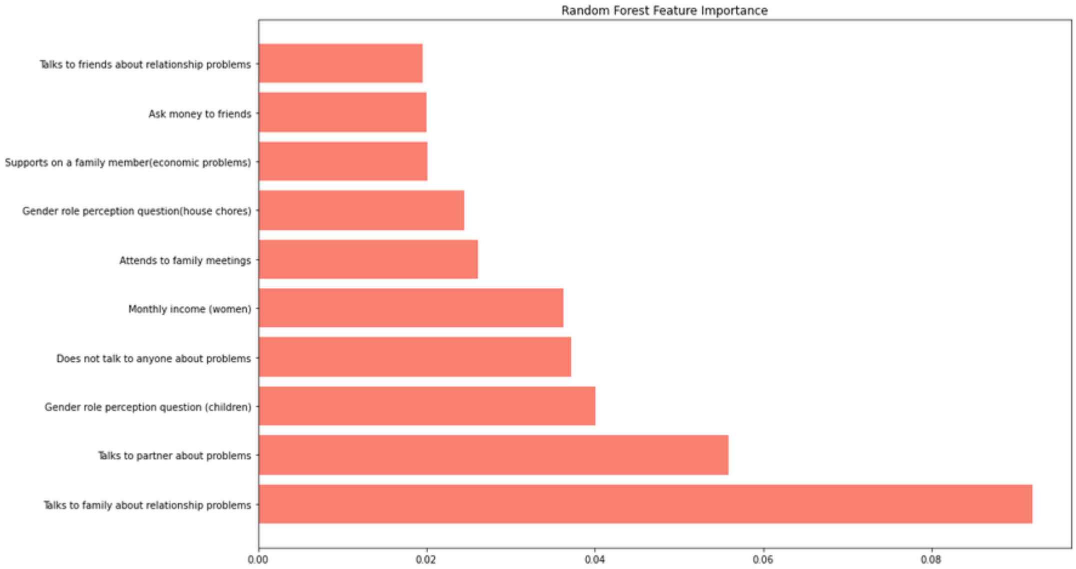


Figure 4: Precision-Recall curves

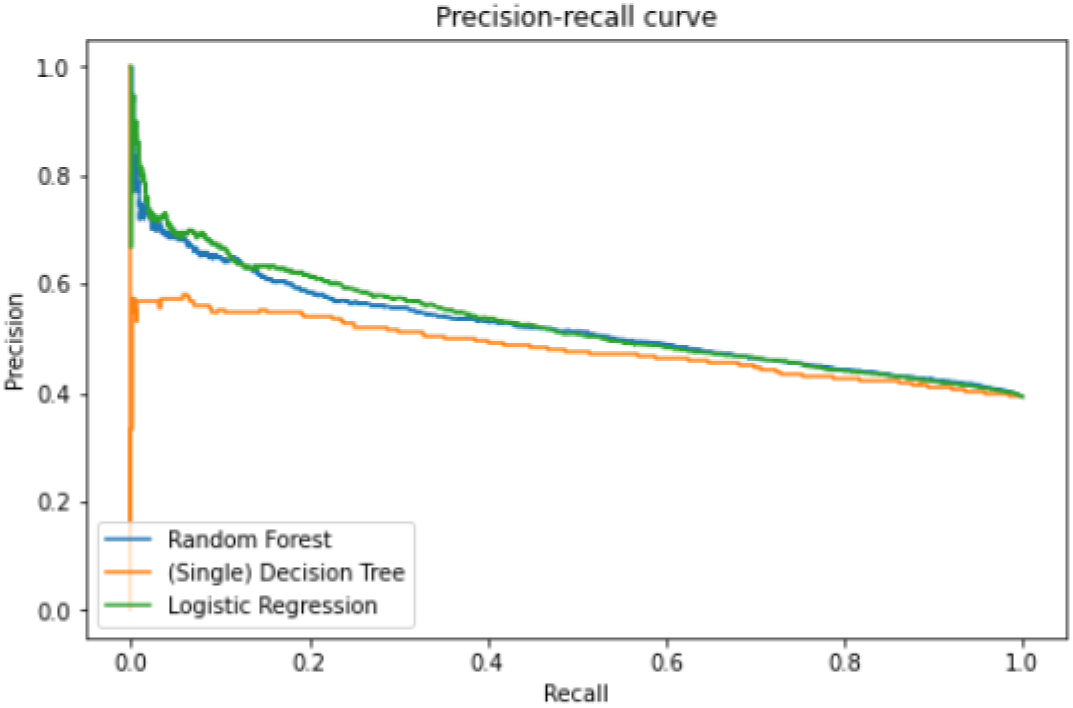


Figure 5: Prevalence of GBV

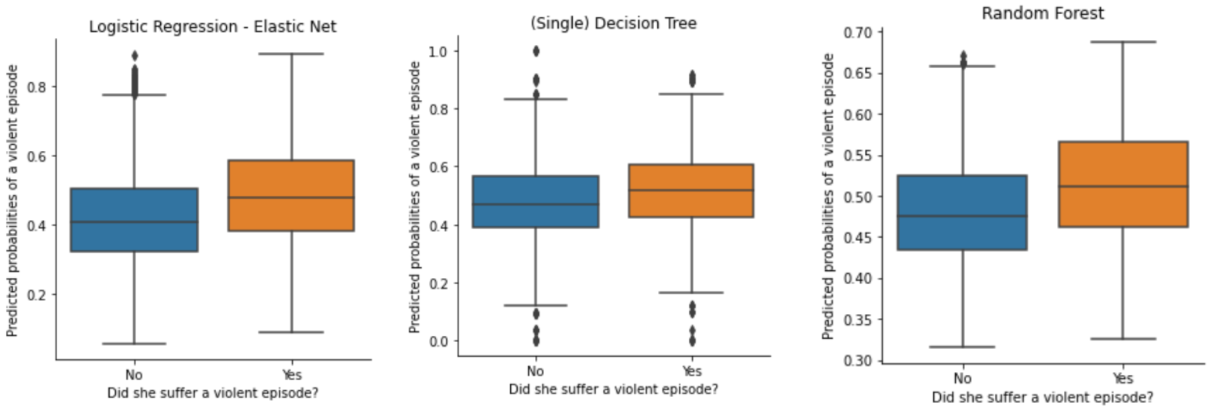
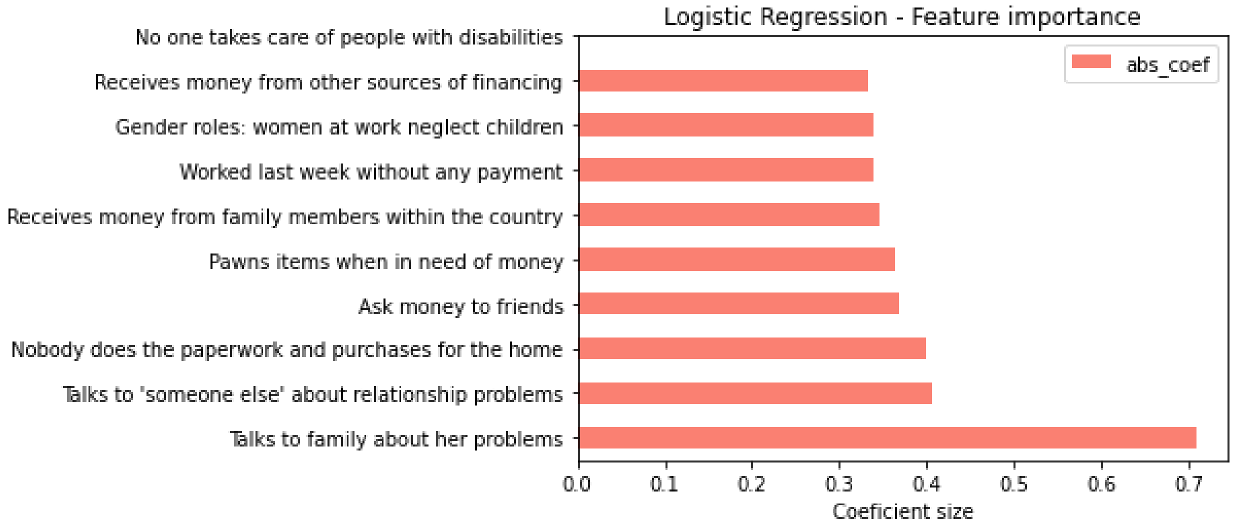


Figure 6: Feature importance of Logistic Regression


Logistic Regression - Feature importance
abs_coef
No one takes care of people with disabilities
Receives money from other sources of financing
Gender roles: women at work neglect children
Worked last week without any payment
Receives money from family members within the country
Pawns items when in need of money
Ask money to friends
Nobody does the paperwork and purchases for the home
Talks to 'someone else' about relationship problems
Talks to family about her problems
Coeficient size