Analysis and prediction of violence against women in Mexico*

1st Esthephany Ayala Yanez˜
School of Engineering and Science
Tecnologico de Monterrey
Monterrey, Mexico
a00818207@itesm.mx

Abstract—Violence against women has been and continues to be one of the clearest manifestations of inequality, subordination and power relations of men over women. In short, women experience violence simply because they are women, and the victims are women of any social stratum, educational level, cultural or economic. Statistics indicate that at least 6 of 10 Mexican women have faced an incident of violence (2017) .There have been previous studies, which analyzed possible violence factors against women, however there are currently no studies on predicting violence against women in Mexico because the open data provided by the institutions are not focused on this issue. The objective of this paper is the analysis and prediction of violence against women in Mexico using the CRISP-DM methodology and Machine Learning algorithms. The data that will be used for this is provided by the"Secretariado Ejecutivo del Sistema Nacional" and the Victims of Crime provided by"Fiscalia General de Justicia" from Mexico City.

Index Terms—Data science, Machine Learning, VAW

I. Introduction

As attention has become more prevalent in the past decade on the persistence of VAW, the need for more and better data to report and address this human rights violations has intensified. Defenders of women and their security want to understand the nature and magnitude of violence. They seek information and guidance on how statistically sound data can be collected on a subject that, though present and often pervasive in our society and culture, is sensitive and often hidden [1].This VAW data is of utmost importance in helping to rate and quantify problems, inform policy, and design evidence-based programs. Perpetrator data and information on locations of incidents of violence can inform prevention efforts and allow further advocacy for policy change. Tracking data over time and tracking trends can help program designers and implementers evaluate the impact of their programs more effectively. It is necessary to know widely with real data the incidence, degree, in which gender violence is permanent in Mexico to better understand the problem. Data is needed to be able to react to the crisis as it deserves.

In this work, we wish to contribute to the analysis of factors associated with women and girls experiencing violence. The aim is to build a model that can effectively predict the likelihood of occurrence of gender crime, depending on spatial and temporal factors; thus, making it possible to prevent them. To achieve this, the study will use machine learning and data science techniques.

However, little research has been found regarding the prediction of gender violence. The potential of machine learning in violence against women and girls forecasting in Mexico is still unexplored and that the power of collected data is still insufficiently exploited.

II. Materials and Methods

A. Data to analyze

As has been seen globally during the development of the pandemic, having timely data and information is crucial to determine strategies and courses of action, as well as to better allocate resources [5]. The same is true for the case of violence against women. In Mexico, official sources provide periodic information on violence experienced by women inside and outside their homes, on which is the data published monthly by the Executive Secretariat of the National Public Security System (SESNSP) [2].

The SESNSP is responsible for setting the basis for coordination in public security in the country. It offers open data resources on common crime incidences. The crime incidence refers to the alleged occurrence of crimes recorded in preliminary investigations or investigation files initiated, reported by the Attorney General's Office of each state [2]. These data are of great value as it contains information such as the type of crime as well as the age range, and modality with which it was carried out by each entity of the country.

Within the data generated by SESNSP, it is possible to consult open criminal investigations for the crimes of domestic violence, homicide and femicide, as well as calls made to the 911 emergency number related to violence against women. The SESNSP data however have several shortcomings, which prevent us from knowing with the necessary degree of detail and speed how violence affects women.

The information shared by SESNSP was the input data for this work. The time frame to study the number of crime incidences shared by SESNSP, was from January 1st, 2015 to December 31st, 2021.

Ano˜	Ranges from 2015 to 2021
Clave Entidad	Contains the id of each state in the country ranging from 1 to 32 .
Subtipo de delito	These are the alleged crimes committed against victims. The only crimes considered for this analysis are intentional homicide and injury, femicide and abduction.
Modalidad	This is how this crime was committed. With firearm, knife, or other.
Rango de Dead	The age range of the victims of the crime. Minors (0–17 years),Seniors (18- older) and unspecified
Sex	Female
Total	The total number of victims of the crime

TABLE I

ATTRIBUTE CRITERIA USED IN SESNSP DATA SET

B. Data preparation

The data collected provided by SESNSP are extensive, therefore, not all of it is relevant to the analysis. A filtering process was performed using the data that are only of interest to us. For the dataset, Table 1 shows the attributes it contains, and the criteria that were used:

A total of 8064 observations were obtained, having an approximate of 1248 observations for each year . To explore the information contained in the country in crimes related to violence against women, including the year, entity, the type of crime, modality in which the crimes were perpetrated, the age range of the victim and the total number of crimes for each entity, modality, type of crime and age range of the victim

C. Data cleaning

The data had 18 different features, both categorical (modality, subtype of crime, age range, etc.) and numerical (victims by month, year, id of entity, etc.). The target variable is the total of victims by type of crime.

The first step is to load the dataset into a dataframe for easy manipulation and exploration. The variables”type of crime” and “sex” were eliminated as they were redundant and did not add extra value to the analysis. In the case of time series analysis, a tidy data process was used because of the structure of the original database.

III. EXPLORATORY ANALYSIS

A. Descriptive Analysis

Once the data were cleaned, the next step was the implementation of a descriptive analysis of the data. Descriptive analysis is a critical step as it generates accessible insights from otherwise uninterpreted data. Due to the structure of the dataset, we used conditional aggregate operations to obtain the results.

In this section, we aimed to determine the total number of victims for each crime (intentional homicide, intentional injury, femicide and kidnapping) for each year from 2015 to 2021. Then, we tried knowing the behavior of each crime over the years. For even more detail, the behavior of each crime, for each month over the years, was explored to find trends within the dataset.

The above process was repeated for the remaining categorical variables (modality and age range) and for each of them, aggregate operations were used to determine their behavior with respect to the subtype of crime over the years, such as the age range of the femicide victims.

B. Time Series Analysis

Time series analysis helps understand the underlying causes of trends or systemic patterns over time. We can see the seasonal trends and dig deeper into why these trends occur. For this analysis, a Time Series Decomposition was performed using Additive model. In addition ARIMA and Random Forest methods were used for time forecasting. ARIMA is a forecasting algorithm based on the idea that the information in the past values of the time series can alone be used to predict the future values. Random Forest is an ensemble learning method and can be used in time series forecasting, both univariate and multivariate datasets by creating lag variables and seasonal component variables manually.

IV. RESULTS

In general results, Figure 1 shows that the year with the highest number of female victims of attacks was 2019, with a total of 70,185 victims. Behavior during the years of Each type of crime remained similar, with intentional injuries being the crime with the highest occurrence during the years, followed by intentional homicides, femicides and kidnapping.

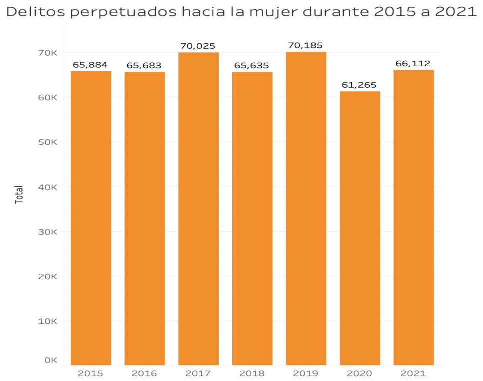

Fig. 1. Crimes committed against women during 2015 to 2021

During the preliminary analysis we can observe that for each crime, the trend is the same, increasing over the years. However, something interesting to highlight was how abductions of women were decreasing, which could mean that now the women victims were no longer being abducted, but were now being murdered.

When comparing femicides with abductions, we can see in Figure 3 how the graphs behave in the opposite way: while the abduction of women was decreasing, femicides were increasing at the same time.

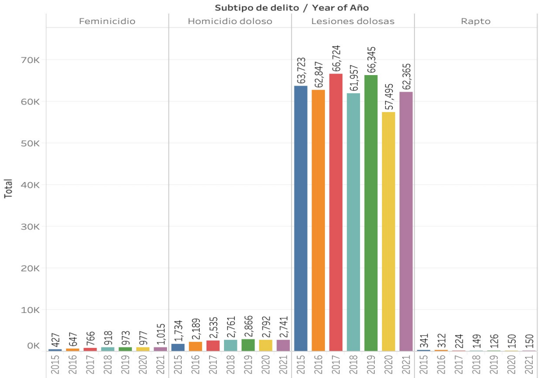

Fig. 2. Crimes committed against women during 2015 to 2021

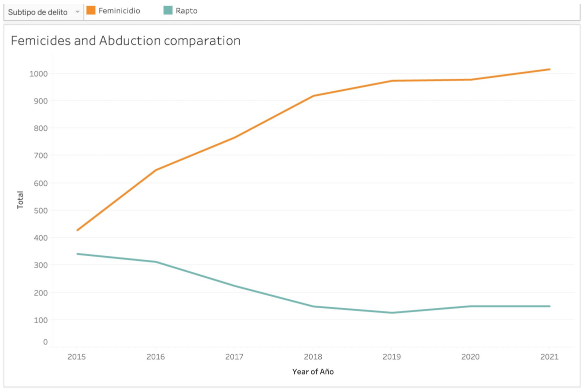

Fig. 3. Comparision of Femicide and Abduction

In the analysis, we could see that the modality of attack over the years was mostly with another element and that the age range of the victims was mostly adult women over 18 years of age. However, for the crime of intentional homicide, the modality of the crime was mostly with a firearm and for the type of crime of abduction, the vast majority of the victims were minors.

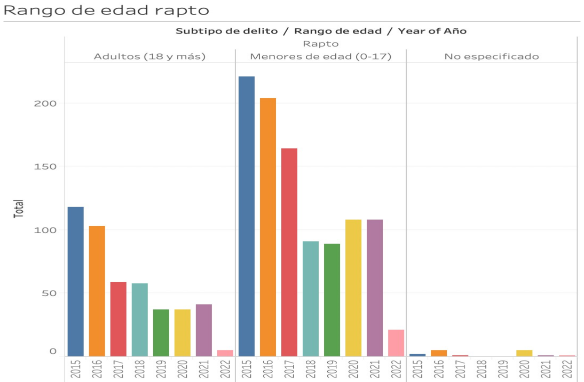

Fig. 4. Age range of abduction victims

Due to the current phenomenon in our country, we decided to focus more on the type of crime of femicide and make our analysis with time series.

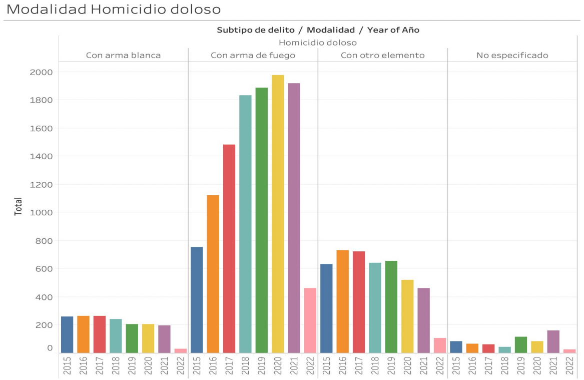

Fig. 5. The modality of intentional homicide.

In a preliminary analysis, we can see the behavior of femicides over the years and say that femicides are increasing and in general features, we can see that it follows a type of trend. To confirm this and visualize our data in a better way, we decided to apply a time series decomposition with an additive model that allows us to decompose our time series into three distinct components, trends, seasonality and noise. Our data confirms that there is a trend and there are peaks in our data and a seasonality of attacks in the spring.

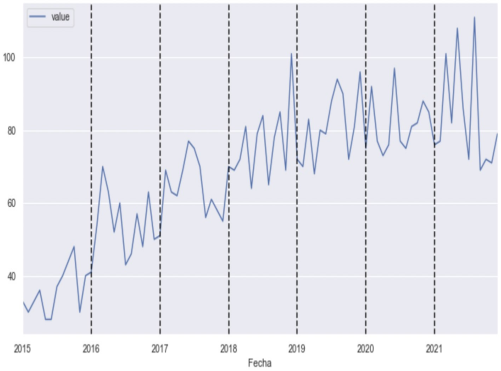

Fig. 6. Femicides Time Series Analysis

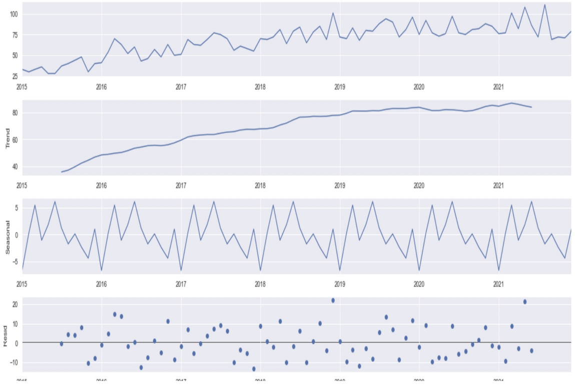

Fig. 7. Femicides Time Series Decomposition

For the time series forecasting of femicides, we decided to use ARIMA and Random Forest. For both models, we used the data ranging from January 2015 to December 2019 as our training data and we used the last two years as our test data. The parameters used for ARIMA were (5,1,2) and Table 2 shows the results for both models, using RMSE as our metric resulting Random Forest as a better fit for our model.

Model	RMSE Metric
ARIMA	12.8303
Random Forest	11.8660

TABLE II

FORECASTING MODELS RESULT

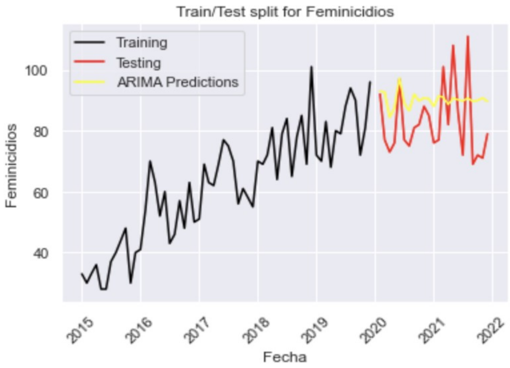

Fig. 8. Time Series Forecasting using ARIMA

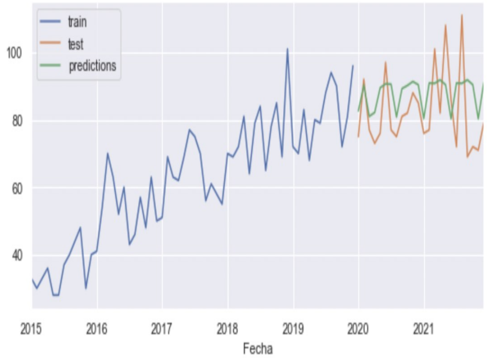

Fig. 9. Time Series Forecasting using Random Forests

V. CONCLUSION

From this work, we can conclude several facts. First, attacks against women continue to grow and in the case of femicides, unfortunately, they will continue to grow. It is interesting to see how only a crime was decreasing but that does not mean a decrease in crimes, but that now these crimes have become major crime resulting in murdered women. Regarding femicides, we can conclude that there is a trend in the years of attacks and a seasonality during the spring, which a reason could be related to the feminist protests in March, however further research is needed on the subject. Further work is also needed on the issue of time series prediction and the use of other models and techniques to improve the results.

REFERENCES

[1] CEPAL. (2020, November). Addressing violence against women and girls during and after the COVID-19 pandemic requires financing, responses, prevention and data compilation.

[2] Secretariado Ejecutivo del Sistema Nacional de Seguridad Publica.´ (2021, April). Incidencia Delictiva del Fuero Comun.´ https://drive.google.com/file/d/1MoF8imFewbL$_1$6*FnlieEA7fNw ZRijTTR/view*

[3] Amusa LB, Bengesai AV, Khan HTA. Predicting the Vulnerability of Women to Intimate Partner Violence in South Africa: Evidence from Tree-based Machine Learning Techniques. J Interpers Violence. 2020 Sep 25:886260520960110. doi:

10.1177/0886260520960110. Epub ahead of print. PMID: 32975474.

[4] Uddin, S., Khan, A., Hossain, M., Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19, Article 281.
https://doi.org/10.1186/s12911-019-1004-8

[5] Equis. (2020). Las dos pandemias: Violencia contra las mujeres en Mexico en el contexto del COVID-19. https://equis.org.mx/wp-´ content/uploads/2020/08/informe-dospandemiasmexico