Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 10, 4251-4276

Exploring The Gender Inequality Gap: Pipelines For Open Source Data Management, A Path To AI Practice

Ana Luna[1], Pilar Hidalgo-León[2], Rafael Ricardo Rentería[3], Andrea Montaño Ramírez[4]

[1] Department of Engineering, Universidad del Pacífico, Peru
[2] Department of Engineering, Universidad del Pacífico, Peru
[3] Observatorio de Demografía y Epidemiología del Área Andina, Universidad Nacional de Colombia, Colombia
[4]Department of Statistics, Universidad Nacional de Colombia, Colombia
[1] ae.lunaa@up.edu.pe; [2] pv.hidalgol@up.edu.pe; [3] rafaricard@gmail.com; [4] apmontanor@unal.edu.co

Abstract— This research article explains in detail the pre-processing stage unifying various techniques, using real and open public data from Peru, between the years 2016-2019. The main objective is to address the study of gender inequality through clean and reliable data. This article shows how to group and clean 6 data sets by category to identify and interpret inequality factors, extract valuable information that can be used in data mining models, and contribute to future decision making. The pre-processing techniques were validated using various prediction algorithms and their performances were compared using ranking metrics.

Keywords— Pre-processing, Gender Inequality, Data Mining, Classifiers, Predictive Algorithms.

I. INTRODUCTION

According to Rubin Gayle, the sex-gender system is the explanation of the dynamics of social inequality in humanity [43]. Rubin stated that all philosophical currents that criticized social inequality and slavery (Marxism, structuralism, and psychoanalysis) lacked a detailed analysis of the impact of gender. In this way, the primary factor that would determine the inequality was omitted. That division of roles is a cause of the gap between men and women, despite being economically poor or rich, the inequalities imposed by roles separate them even in the same social stratum. In this context, gender inequality should be treated as a priority due to its cascading effect on society and inequality in general. As a fundamental right, gender equality is included in the2030 Agenda of the United Nations Development Program. Its main objective is to have the same opportunities, rights, and obligations in all spheres of life, and to reduce gender inequalities as a symbol of progress and social improvement [27], [12], [40]. For this, the importance of using technology to achieve long-term goals is required, and Computer Science techniques are undoubtedly an essential tool.

The Inter-American Development Bank (IDB) has also taken an interest in this perspective, creating social indicators that require the development of women; and their rights are priority components in the economy of the nations. Initially, efforts to study gender inequality focused on immediate origins, such as home conditions and domestic environment [36]. However, underlying factors (values, attitudes, and cultural traditions) that determine, political and economical possibilities for women and men were omitted [1], [2], [3], [23], [47]. In recent years, contextual factors explain the latency of inequality and its consequences have been gradually incorporated in studies [5], [33], [34], [22], [48]. Different gender indexes have been proposed in order to have a tool that quantifies and measures the level of inequality. The particular case of the Gender Inequality Index (GII) is usually used in developed countries by researching the following domains: Health, Empowerment, and the

Labor Market. But, considering the features of each country, it could be employed for every region [31]. Gender inequality must be treated as a priority due to its cascading effect on society and inequality in general. Peru's national surveys showed worrying data on the statistics of femicide and mistreatment where more than 120 women were murdered in 2017 [42]. On the one hand, in recent years progress has been made in the regulatory framework given by the Peruvian Congress, such as the law on equal pay for men and women; but, in legislative matters, much remains to be done compared to other countries [32].

The first step to reduce inequalities is to have a real diagnosis through gender-sensitive indicators. The importance of gender indicators is mainly because they make visible what was hidden and improve the evolution of programs and projects that contemplate different areas of social life such as politics, economics, education, and rights; where the magnitude of inequality is considerable. The system of gender indicators allows to have an initial panorama within a global context, it also contributes to the monitoring and modification of the growth of inequality [35], [4], [19], and to the development of public policy projects to reduce gender inequalities [37], [44], [60]. In this research work, we gathered and described the process for the construction of six heterogeneous public data sets, corresponding to the Peruvian population during the years 2016, 2017, 2018, and 2019 for the study of gender inequality after a rigorous methodology to obtain clean data sets to be used in data mining models. This information can be employed for a wide variety of applications, such as the GII estimation, correlation studies, recognizing inequality factors in education, vocational skills, economic resources and opportunities, and social norms that promote gender equality by identifying the negative combination of cultural patterns and institutional limitations. The key to a successful study requires, previously, the provision of data and statistics disaggregated by sex and age, and it also takes into account the different roles, tasks, and responsibilities of men and women in society, to analyze gender inequalities without biased values. The way of the acquisition of the original data is the core to not condition a post-processing trend. The article is structured as follows. The following section shows the details of the literature review addressed during this work. In section 3, we describe the data collected; as well as the methodology used in the pre-processing of each data set depending on their imperfections. Then, in section 4, we present the results after the pre-processing, and in section 5, we validate the techniques employed in this work through some possible applications. Finally, the conclusions are shown.

II. LITERATURE REVIEW

The inequality relationship is reflected in all areas. The data is no exception to an inequality bias, the authors of Ref. [13] analyzed and concluded that the goal of making progress on gender equality is related to closing the gender data gap. If the data does not accurately specify the reality of men and women, it will not be possible to satisfy the needs due to the absence of indicators. UN Women called for the collection of sex-disaggregated data to ensure that the needs of women and girls are visible [54]. Those indicators should be based on principles of gender equality and human rights, and then generate quantitative metrics to finally arrive at the inferential statistics process. But what it conclusively requires is that the data reflect the quantitative scope of gender violence that does not occur.

The evolution of gender studies and reflection from the academy proposed gradually incorporated contextual factors that explain the latency of inequality and its consequences [5], [48], [22]. In the particular case of Latin America, international organizations have used this approach to formulate action guidelines and improvement strategies in the fight against inequality (Center for Research and Development (CIDE) 2002). Such is the case of the "Ecological model for a life free of gender violence" [34], which is a holistic perspective that seeks to understand the problem from four areas or factors: 1) Individual factors (men, women, and children who live in inequality and exercise the right of access to education, work, economic independence, etc.) that involves all

members of the family and not just women; 2) Family factors (the tolerance that this closed population has in the face of acts of violence); 3) Community factors (indicators of inequality such as illiteracy, child malnutrition, access to housing, etc.); and 4) Social factors (intervention or lack of it by the authorities, police contingents, courts, etc.). The contributions of these studies have a great impact on addressing gender inequality, however, the methodology for data collection and its pre-processing that contributes to each of the factors mentioned were not cited.

According to some researchers, the rates of gender inequality arise from the comparison between the levels of achievement for women and men in different dimensions of well-being. Its calculation requires specific methods, particularly in multidimensional contexts. The author of the Ref. [39] developed different indicators to measure gender inequality, evaluating its consistency. One of them measures the extent to which one sex can be said to be better than the other. The other index measures the average disparity/difference between women and men taking into account to what extent the general gender inequality favors both sexes equally or not. He also defined an indicator that measures the average amount of gender inequality, regardless of the sex that benefits from it. Finally, he redefined this index to make it sensitive to the extent that existing levels of gender inequality are due to gender gaps that favor only one sex or the other.

In some countries, for example, women perform better than men in life expectancy and education, but not in average income. This means that there is a certain balance between dimensions. For this reason, aspects of measuring multidimensional levels of gender inequality in a given society should be studied. Otherwise, the differences in choosing one indicator over another could be very important, creating a non-negligible problem of compensation between dimensions

The focus of this research work is the construction of indicators and indices as well as the importance of their choice to measure gender inequality. However, the way that the cleanliness of the data collected to address the study is not detailed.

Although the contributions to make inequality visible in different sectors is extremely important to confront this problem, the treatment of the data is the preliminary and fundamental step to reach a deep analysis and generate inclusion measures.

There are a variety of studies in which the qualitative and quantitative data needed for a gender analysis are described [21], [24],[38], as well as research works and manuals for the identification of gender inequalities through the construction of indicators [18], [9], [4]. Although it is known that the gender variable is not enough to show situations or population subgroups where inequities are more evident, it is essential that at the time of the design of the data capture instrument the variables to be recorded and the pre-processing to follow must be clear.

In a recent UNESCO report (UNESCO, 2020), the main findings of the contributions of experts in Gender Equality and AI (Artificial Intelligence) are disclosed. They recommend addressing gender equality considerations through AI, which involves using computers to classify, analyze, and make pre-dictions from data sets, using a set of rules called algorithms. The main objectives are to identify patterns, make predictions, and recommend actions to increase the visibility of this phenomenon. The report highlights that AI systems are potential tools that would actively correct gender inequalities providing a great opportunity. Along the same lines, a report published by the University of Cambridge describes some of the most important challenges for the study of gender equality presented by recent developments in artificial intelligence (AI) [8]. One of them is biased data sets that are often not representative of the audience demographic. Such types of data sets amplify racial and gender inequality. For this reason, the collection, management, and purpose of large data sets need to be further explored, including gender and context-specific guidelines for best practices for data.

For the particular case of Peru, and according to the National Institute of Statistics and Informatics in 2017, the gender inequality index (IGI) [32], showed that less than 50% of women had a job, while in men, this percentage reached 83%. Regarding the educational aspect, the same indicator revealed that 74% of the men finished secondary education. On the other hand, women do not reach 63%. In the same sense, in the field of political participation and senior management positions, women did not exceed 25%. Also, national surveys showed worrying data on the statistics of femicide and abuse where more than 120 women were murdered in 2017 [42]. Finally, in recent years, progress has been made in the regulatory framework given by the Peruvian Congress, such as the equal pay law for men and women; but in legislative matters, there is still a lot to do compared to other countries [32]. Although there are enough open data sources to carry out studies of gender inequality, their treatment has not been studied or unified in depth so far.

In this work, we use open data and describe in detail the pre-processing methodology of various open data sets that optimize the study of gender inequality in Peru in various areas, such as public offices in the three branches of Congress, access to paid work, and education, among others. Data collecting and pre-processing stages with appropriate techniques that minimize bias for subsequent decision-making or implementation of a prediction model are crucial and will be detailed in the next section.

III. METHODOLOGY

Through holistic data, it is possible to analyze the conditions of inequality and gender-based violence. The data showed in this work is freely available and the information is at the departmental level, within Peru during the study period from 2016 to 2019 (The 6 Data Batches are available in the GitHub repository: https://github.com/IngenieriaUP/EGENDERIN-EQUALITY-GAP. In the same link, there are also 6 open-source notebooks written in the Python programming language that include the backbone of the pre-processing stage of this work). This section specifies the steps to generate clean data sets from the original ones in the Peruvian context. The data collected and analyzed correspond to the following variables: access to education, paid work, political representation, social relations, teenage pregnancy, maternal mortality, domestic violence, records of police attention for crimes associated with gender violence, information from the media on femicide and sexual abuse.

According to [45], [16], the first step to obtaining efficiency and good performance from the models applied in data mining is the optimization of the pre-processing stage. Also, this is the first step of the KDD or Knowledge Data Discovery process, proposed by [11], subsequently modified by [53]. There are different techniques to obtain valuable patterns and insights from data sets [59], and getting relevant information is crucial and should add efficiency to the KDD process. For [46] and [14], data collection (within a data set) can lead to availability problems. One way is the manual collection, which is the case of the data available in repositories (eg: tables on State web pages). The second way is through a web scraping technique and it is usually used when the data is not contained in a database or repository (eg: news, tweets, etc.). This last technique is applied through crawlers to extract information from websites, avoid errors in obtaining data, and automate this process. Sometimes these crawlers simulate human beings browsing the web, either by using the HTTP protocol manually or by embedding a browser into a paraphrase application (The controversy over the legality of Web scraping is resolved when the sources are publicly accessible or the data is collected for a purpose of general public interest).

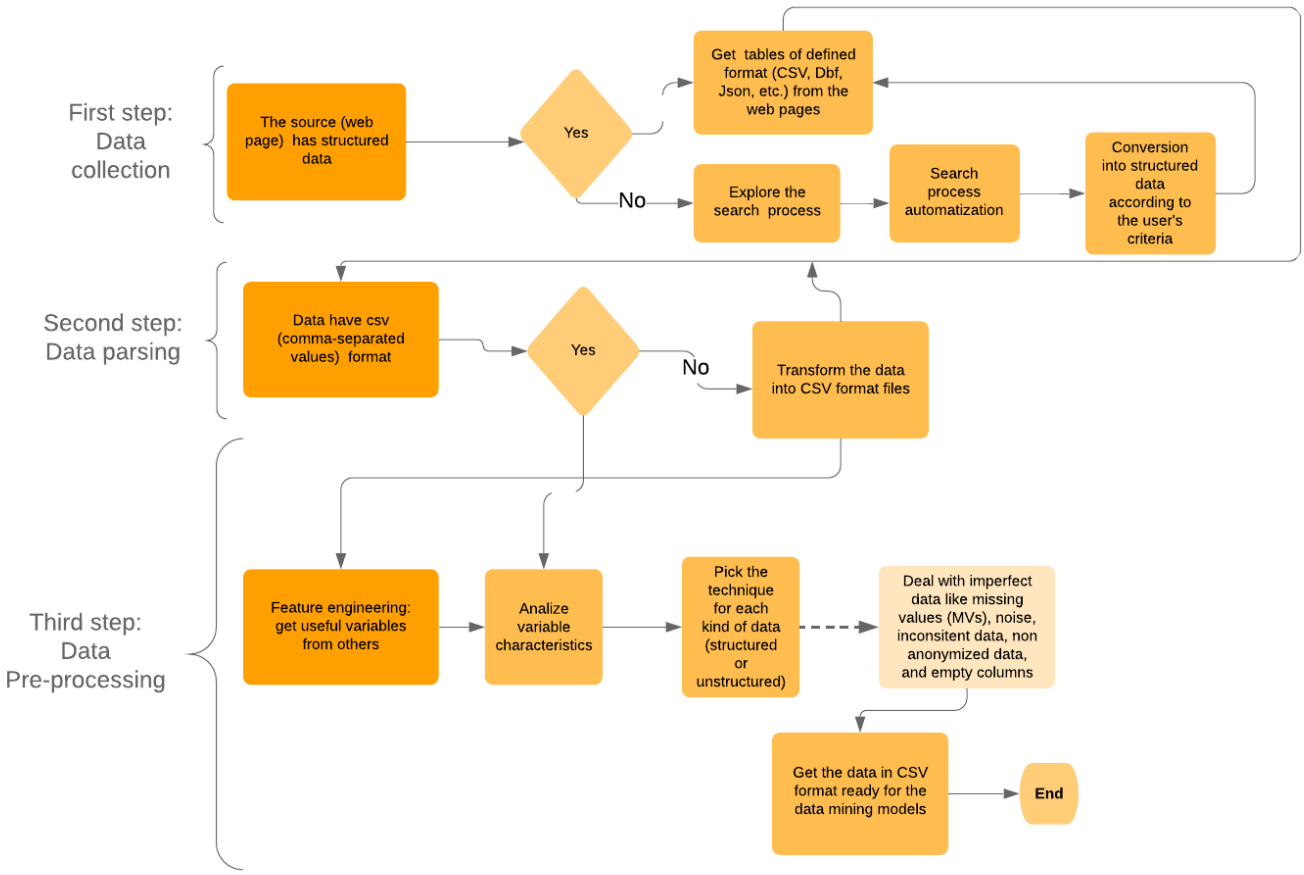


Fig. 1: Different steps of the applied KDD process

In this research work, we followed the process of collection, exploration, and pre-processing of the data set described in Fig. 1. Next, we detailed the process of obtaining information through a series of resources, such as syntax analysis or data parser (analyzer), which transformed updated data into structured sets that could be pre-processed through the feature engineering process: Data filtering (variables, objects, and instances) and cleaning (inconsistent data, noise, and missing values). In the following subsections, we described each of the stages outlined in Fig. 1.

A. First Step: Data Collection

As described in Fig. 1, in this work the data collection was carried out following two paths depending on the type of information, structured or unstructured data. For the first one, the repository available online has the location and specific links for each data set within transparency portal of the Peruvian state (Available in www.transparencia.gob.pe/) and national surveys. Next, the information contained in each of the structured data sets is described:

1) ENDES: It is the Demographic and Family Health Survey and has been carried out annually since 2000 by the National Institute of Statistics and Informatics (INEI).

2) ENARES: It is the National Survey on Social Relations which is carried out every 4 years by the INEI.

3) Legislative authority: These data are from the workers' payroll of the Congress of the Republic of Peru, they have been openly published monthly since 2015.

4) Executive authority: These data are from the workers' payroll of the Presidency of the Republic of Peru, they have been openly published monthly since 2013.

5) Judicial authority: These data are from the workers' payroll of the judiciary of Peru, they have been openly published monthly since 2017.

6) Police Complains: It is the annual registry of crimes in the police stations of Peru.

In Table 1 the structured data collection's characteristics are detailed as well as the variables measured in one of them for the study of gender inequality (fifth column). Finally, data batches are also described.

Once the files from each repository were downloaded, they were grouped by year, e.g. for the dataset of the Executive Authority, 12 files correspond to the months of the year, however, it was recommended to collect them all in one file to have a complete overview of what happened in that period. Some data sets belong to surveys with multiple sections, such as the ENDES (13 sections) and ENARES (15-22 sections) surveys depending on the different parts of it. As a consequence, the categories were grouped according to the code of the respondent, preserving the variables that must be considered for the gender inequality study [41], [17], [34],[4]. Then, structured data were merged and labeled in a batch to be pre-processed.

For the second type of data, the unstructured ones or the data that was not in tables, the process was different. In the case of News, the data set was obtained using the scrapping technique with an application developed in Python 3.7 and the newspaper package that allowed us to navigate through each page searching keywords and obtaining a set of data such as their link, title, content and publication date. The format for storing all these data was in "CSV". The generic and specific description of this unstructured data is outlined below and in Table 2, respectively:

TABLE I
OPEN STRUCTURED DATA SETS OF OFFICIAL WEBSITES OF PERU AND NATIONAL SURVEYS FROM 2016 TO 2019

Source	**Files**	**Year**	**Size**	**Variables to Measure Inequality**	**Data Batch**
ENDES	16	2016	4,8 GB	Maternal mortality, teenage pregnancy, labor force and education	ENDES 2016
	16	2017	4,8 GB		ENDES 2017
	16	2018	4,8 GB		ENDES 2018
	16	2019	4,8 GB		ENDES 2019
ENARES	15	2015	3,7 GB	Social relations and gender roles	ENARES 2015
	22	2019	4,0 GB		ENARES 2019
Legislative authority	12	2016	7,5 MB	Women empowerment	Legislative 2016
	12	2017	7,5 MB		Legislative 2017
	12	2018	7,5 MB		Legislative 2018
	12	2019	7,5 MB		Legislative 2019
Executive authority	12	2016	18,3 MB	Women empowerment	Executive 2016
	12	2017	18,3 MB		Executive 2017
	12	2018	18,3 MB		Executive 2018

	12	2019	18,3 MB		Executive 2019
Judicial authority	12	2017	32,4 MB	Women empowerment	Judicial 2017
	12	2018	32,4 MB		Judicial 2018
	12	2019	32,4 MB		Judicial 2019
Police Complaints	1	2016	74 MB	Violence attending	Complaints 2016
	1	2017	99 MB		Complaints 2017
	1	2018	164 MB		Complaints 2018
	1	2019	138 MB		Complaints 2019

TABLE II
DATA SETS BUILT FROM 2 NEWSPAPER SITES IN PERU FROM 2016 TO 2019. THE 2 KEYWORDS SEARCHED FOR THE STUDY OF GENDER VIOLENCE WERE: "SEXUAL ABUSE" AND "FEMICIDE"

Source	**Files**	**Year**	**Size**	**Variables to Measure Inequality**	**Batch Data**
News paper El Comercio	1	2016	0,7 MB	Feminicide attending	newsCfem2016
	1	2017	1 MB		newsCfem2017
	1	2018	1 MB		newsCfem2018
	1	2019	4 MB		newsCfem2019
News paper La República	1	2016	1,2 MB	Feminicide attending	newsRfem2016
	1	2017	1 MB		newsRfem2017
	1	2018	3 MB		newsRfem2018
	1	2019	2 MB		newsRfem2019
News paper El Comercio	1	2016	1 MB	Sexual Violence attending	newsCabs2016
	1	2017	2 MB		newsCabsfem2017
	1	2018	2,2 MB		newsCabsfem2018
	1	2019	5 MB		newsCabsfem2019
News paper La República	1	2016	1,4 MB	Sexual Violence attending	newsRabs2016
	1	2017	1,3 MB		newsRabsfem2017
	1	2018	2 MB		newsRabsfem2018
	1	2019	2,5 MB		newsRabsfem2019

1) News: This data set belongs to the exploration of 2 web pages of Peruvian newspapers from 2016 to 2019. Specifically, 2 different keywords were searched: "Sexual

abuse" and "Femicide", both classified as crimes in Peruvian criminal law and within acts of gender violence according to [34]. (See Table 2).

One of the characteristics of unstructured data is that it is very rich in content. However, it is necessary to preprocess unstructured information to allow its incorporation into general problem solving or decision-making processes, according to the user's needs. In this context, the best decision is to clean the data set while preserving the integrity of the text.

The next stage, which is described in the next subsection, consisted of grouping all the data collected for the study of gender inequality in the same format to ensure that the pre-processing stage is developed uniformly.

B. Second step: Data Parsing

The data parsing process is when one data format is transformed into another and it becomes more readable. A parser takes the data in a particular format and transforms it into a more readable data format that can be easily read and understood, e.g. if a *.txt plain text file does not contain a column break, the variable data cannot be distinguished. A well-made parser will distinguish which information is needed, and following the parser's pre-written code and rules, it will pick out the necessary information and convert it into CSV. If this process does not carry out, the pre-processing stage is not possible, unless certain criteria are established to interpret the information they contain. Table 3 shows the identified files that have different formats in each dataset, which proves that it is necessary to unify the files' format in order to obtain datasets that can be readily used in a stage of training algorithms or by the same processing tool.

The final format to homogenize the data sets was the "CSV" that separates the columns in a table by commas and each record on a single line. After this process, the dataset is ready for filtering and cleaning techniques.

TABLE III
FILE EXTENSIONS OF THE COLLECTED DATA SETS

Batch Data	**Files**	**Original Format**	**Parse Format**
ENDES2016-2019	9	DBF	CSV
ENARES2016,2019	13	SPSS	CSV
Legislative2016-2019	1	.xlsx	CSV
Executive2016-2019	1	.xlsx	CSV
Judicial2017-2019	1	.xlsx	CSV
Complaints2016-2019	1	.xlsx	CSV
newsCfem2016-2019	1	HTML	CSV
newsRfem2016-2019	1	HTML	CSV
newsCabs2016, newsCabsfem2017-2019	1	HTML	CSV
newsRabs2016, newsRabsfem2017-2019	1	HTML	CSV

C. Third step: Data pre-processing

Real-world data is often incomplete, inconsistent, lacking in certain behaviors or trends, and is likely to contain many errors. To tackle this problem, the data pre-processing technique is introduced. The data pre-processing stage is the set of techniques (data preparation and reduction) used before the application of a data mining method [14], to adapt the data to the requirements posed by each mining algorithm data, allowing the processing stage [55], [16]. However,

pre-processing techniques have to deal with imperfect data, so it is necessary to validate its consistency. For example, missing values (MVs) and noise data could affect the performance of the models [56], to avoid this, it will require the application of various methodologies to ensure the precision and reliability of the data.

At this stage, and taking into account the focus of the role within the study context, gender inequality, we use the process of feature engineering. This is a technique to extract or generate data variables and/or characteristics, such as some hidden knowledge from the raw data, and convert them into some important characteristics. Finding attributes makes it easier to understand, in the context of a problem, the characteristics that are important to predictive models.

The objective in this phase was to follow a series of steps necessary to obtain clean and useful data sets. Each data set went through two stages: first, analyzing the characteristics of the data by identifying MVs and inconsistencies; and secondly, the data imputation stage, where the imperfect values identified in the first phase change, generating variables with clear values, for example by replacing the missing numerical values with the average value or building a new column "gender" based on the names of the people in the data set [57].

In this step, structured and unstructured data must be differentiated for pre-processing and the techniques to be used.

1) Structured Data Analysis: The first approach in this stage was the use of descriptive statistics to analyze and characterize each of the collected data sets. We first evaluated whether the data values followed a normal distribution. Also, we calculated the percentage of MVs for each data set. Table 4 details the "Type of imperfection" that we identified for each data set, we showed examples; and specified the choice of the strategy carried out to correct the defect, according to the bibliographic review.

The strategies carried out for each type of imperfection are detailed and developed in the subsequent subsections.

Missing Values: Some data sets in ENARES and ENDES Batch had MVs, their percentage was evaluated and the columns with 100% MV records were removed. Another case was the one in which some variables only recorded one data for all the observations, such as the "Year" field in the data set

"WOCAP1002019" that corresponds only to the value "2019". The presence of MVs affects the data analysis since it is assumed that the data sets follow a normal distribution [52]. Figure 2(a) shows that, before cleaning the MVs of the data in the "WOCAP1002019" dataset, the RFINAL variable does not correlate with the other variables. However, Fig. 2(b) presents how the distribution and correlation between "WOCAP1002019" dataset's variables are preserved after the cleaning process.

For the cases where the imperfection type corresponds to MVs and applies the data imputation technique, the defect was replaced by the mode or the mean. The imputation process consists of replacing the MVs with the most representative data of the sample. MVs are the ones that have not been registered, for any reason, like not being recorded by the team during the survey or not having evidence that warrants the registration, among others. One of the ways to deal with these problems is to discard values. However, there is a high risk of eliminating important information, especially if the MVs percentage is greater than 10%, which could affect the inference process in the implementation of models of data mining, as the appearance of MVs can rarely be ignored. In that sense, the authors of Ref. [15] suggest the use of two methods to

deal with MVs. The first one is a traditional solution, it comes from statistics and it is based on the use of the maximum likelihood procedures that seek to find the most probable values of the distribution parameters for the set of data under study to be used in the imputation stage. The second method proposed by the authors is to use a simple imputation technique like mean substitution. The process of imputation or change of values can alter the distribution of the data. In this work, when the original numerical data had a normal distribution, we used the mean value. However, if the distribution was not symmetric and skewed (positive or negative), the median value was employed, and for non-numerical values, the mode value was assigned [20].

Data Inconsistency: In this type of imperfections, the missing, empty, or "not accurate" values in the analyzed data set correspond to the numbers 99, 999, 9999, 99999, 9998, 998, 98, 9996, 9997. Therefore, when an analysis is evaluated and performed statistically, the values give wrong results.

To solve this problem, these values are replaced by NaN. Finally, with the help of the Pandas statistical package, each data frame was transformed into a "CSV" (comma-separated values) file.

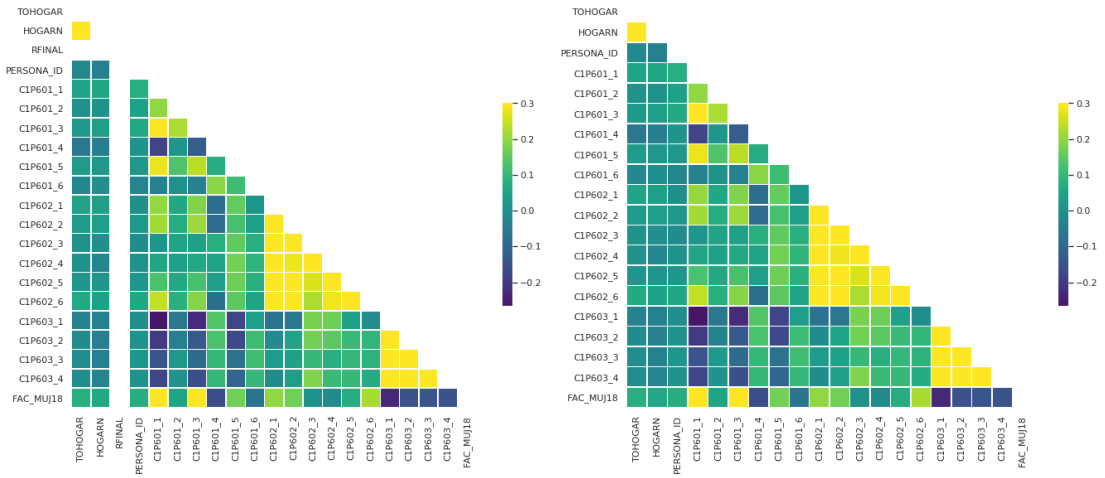
According to the INEI dictionary, values such as "9999" are associated with an MV and correspond to a response not given by the person who has been surveyed. However, in the analysis stage, in many cases, these values are identified as an inconsistency in the records, but through a manual scan, each variable must be carefully studied.

For the Complaint dataset, some of the problems identified were that the numerical data, for example, "AGE" was in a character string format. The solution was to convert it to an integer in a consistent format. Another reported drawback was that the "SEX" column had more than 2 values, but since there were only 4 observations, the dataset was removed.

TABLE IV
IMPERFECTION TYPES OF DATA BATCHES DESCRIPTION. ALL OF THEM WERE COLLECTED FROM 2016 TO 2019. SOME EXAMPLES ARE SHOWN AND THE STRATEGIES TO FOLLOW ARE PRESENTED

Type of Imperfection	Data	Example	Strategy
Missing Values	ENDES, ENARES	Columns with more than 5% of MVs	Column from multiple choice questions in the survey, should be joined and merged
		Columns with less than 5% of MVs	Impute data with the mode or the mean value
		Empty columns	Delete empty columns
Inconsistency in data	ENDES, ENARES, Complaints, Judicial, Legislative, Executive Authorities	Numeric string	Convert string type to integer
		MV with some random value, e.g. 9999 (e.g. Age or Weight)	Replace inconsistent values with'NaN', and then apply an MV strategy
Non anonymized data	Complaints, Judicial, Legislative,	Names	Inferring gender by name
		Telephone, ID	Delete these columns

	Executive Authorities		
Series Name Columns	ENDES, ENARES, Complaints	*IH*2111, *IH*2112..., *IH*2114	Column from multiple choice questions in the survey, should be joined and merged



(a) (b)

Fig. 2: (a) Correlation matrix for the variables of the WOCAP1002019 dataset: The range of colors represents the trend in the negative (blue) and positive (yellow) Bell curve. The 20 variables of the analyzed data set follow a normal distribution; the green squares indicate that they tend towards a central concentration, the yellow one indicates the negative trend and the blue one the positive one. From the figure, we can deduce that by having more green squares, the trend is positive-central with a negative tail on the left. (b) By removing the RFINAL column whose content was the field "Year" with the unique value of "2019", the distribution remains without altering the correlation between variables.

Non anonymized data: For the dataset corresponding to the Legislative, Executive, and Judicial authorities, the imputation process consisted of generating a new variable: "gender" (inspired by https://ialab.com.ar/portfolio-items/i-map/). From the names of the workers' staff, the gender of the people was determined, allowing to know the percentage of men and women working in the State offices. This process is complex and involves the use of Natural Language Processing (NLP) and classification algorithms and exceeds the scope of this work (see the references [6] and [30] for more details). The Python NLTK (Natural Language toolkit) NLP package provides a list of male and female names (500 names in total) to train a classifier to predict whether a name that is not on that list is female or male. To improve the performance of the qualifier, the list of names was fed with the names available at:

http://buenosaires.gob.ar/areas/registrocivil/nombres/busqueda/buscador_nombres.php?&menu_id=16082

Obtaining 6803 female names and 7560 male names. Likewise, the NLTK's classifier is based on two linear prediction models, Naïve Bayes [29] and Decision Trees [25] which are evaluated according to the accuracy, recall, precision, and f-measure classification metrics. The training and test

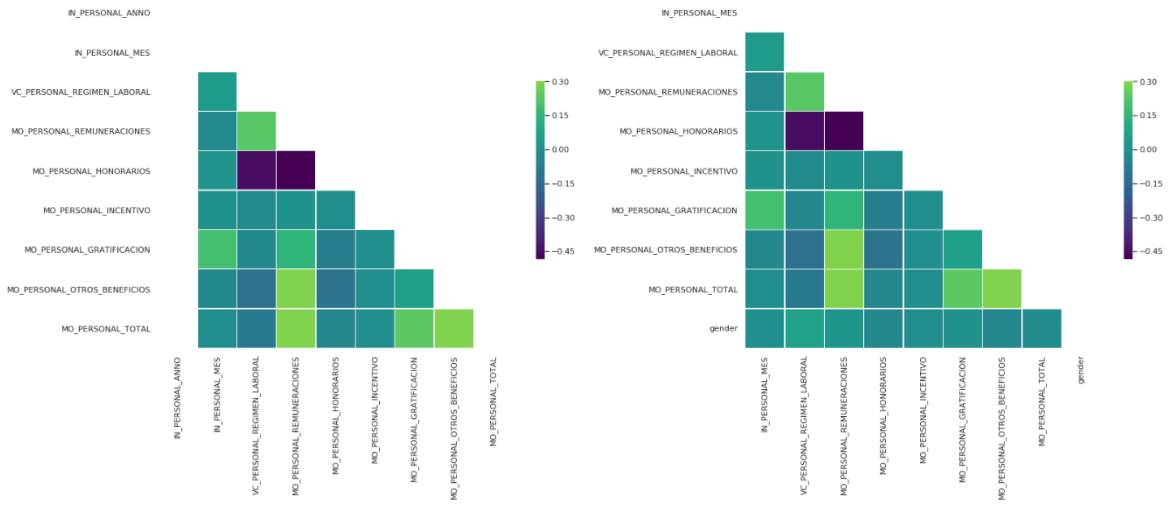
sample are the stages of this process; for the training stage 30% of the list of labeled names was used and 70% for the test sample stage. Table 5 shows the results after applying the classifiers. Both obtained metrics that were very similar to each other, the shorter computational execution time provided by Naive Bayes was privileged. Finally, the values predicted by the classifier were assigned to the 'Gender' column.

Then, for the imputation by MVs, the mode was used for categorical data and the median for numerical data. According to Fig. 3(a), the distribution of the data is left-skewed (tends towards blue), so we use the median for being more representative than the mean. Figure 3(b) shows that the distribution has not changed and the variables with 100% of MVs have been eliminated.

Like the previous data sets, the data imputation was carried out following the criteria of allocating through the mode and the median for numerical data, since after the process the distribution remains unchanged.

TABLE V
METRICS' COMPARISON BETWEEN NAIVE BAYES AND DECISION TREES LINEAR PREDICTION MODELS (NLTK PACKAGE USED FOR GENDER ASSIGNMENT ACCORDING TO A PERSON'S NAME, AVAILABLE AT HTTPS://WWW.NLTK.ORG/BOOK/CH05.HTML)

Classifier	Precision	Recall	F1-score	Accuracy	Execution time (sec.)
Naive Bayes	0,88	0,8	0,79	0,88	1,32
Decision Tree	0,87	0,8	0,79	0,88	3,04



(a) (b)

Fig. 3: (a) Correlation matrix between the variables of the data set Executive Authority, 2019. (b) Matrix obtained after the imputation process.

Series Name Columns: Some of the multiple-choice questions, with matrix structure, were concatenated to give a single answer, and the corresponding label replaced the assigned number. For that purpose,

each "1" was replaced in the matrix by a prime number assigned to each column, as can be seen in the following example:

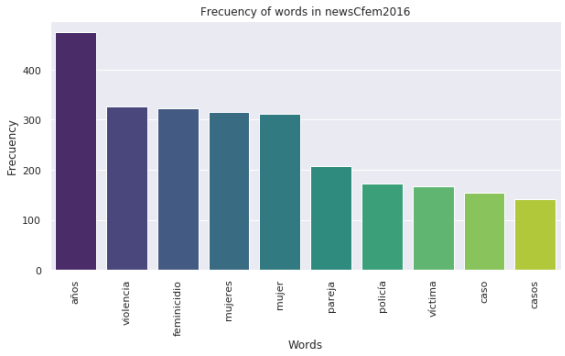
Complaints2016_['IH211_2']=Complaints2016_['IH211_2'].replace['1'],1)

Complaints2016_['IH211_3']=Complaints2016_['IH211_3'].replace['1'],2)

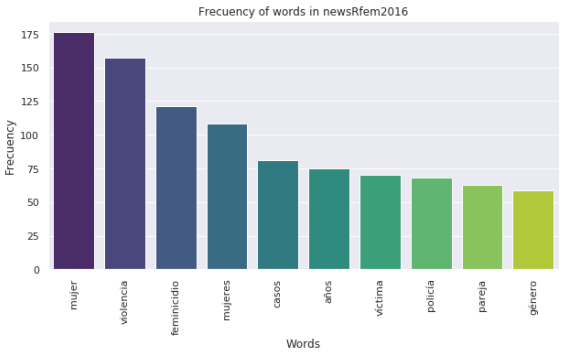
Complaints2016_['IH211_4']=Complaints2016_['IH211_4'].replace['1'],3)

Complaints2016_['IH211_5']=Complaints2016_['IH211_5'].replace['1'],5)

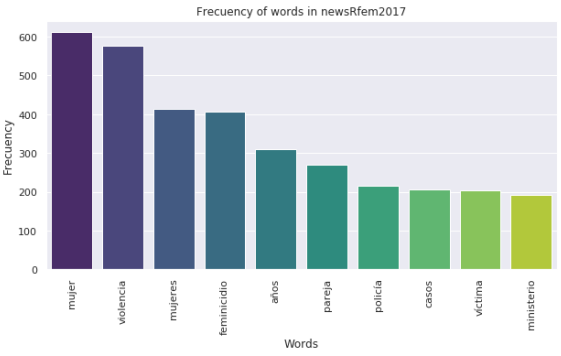
As shown in Table 4, this strategy was applied in the case of Complaints2017 in the column "IH211_4", the values were replaced by the number "3" of those records that contained "1" which symbolizes "committed crime" and the "0" symbolizes "a crime not committed". Therefore, all the data corresponding to crimes has been grouped, and if someone added more columns of a particular registry, that type of crime would be identified, avoiding in this way, replicate columns with the same information. Thus, all the columns that add "3" will have the label that corresponds to the values of "IH211_4".



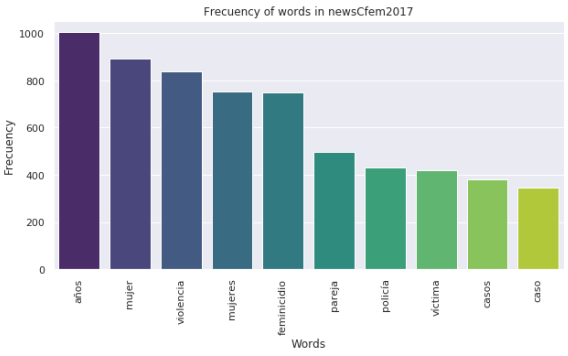
(a)



(b)



(c)



(d)

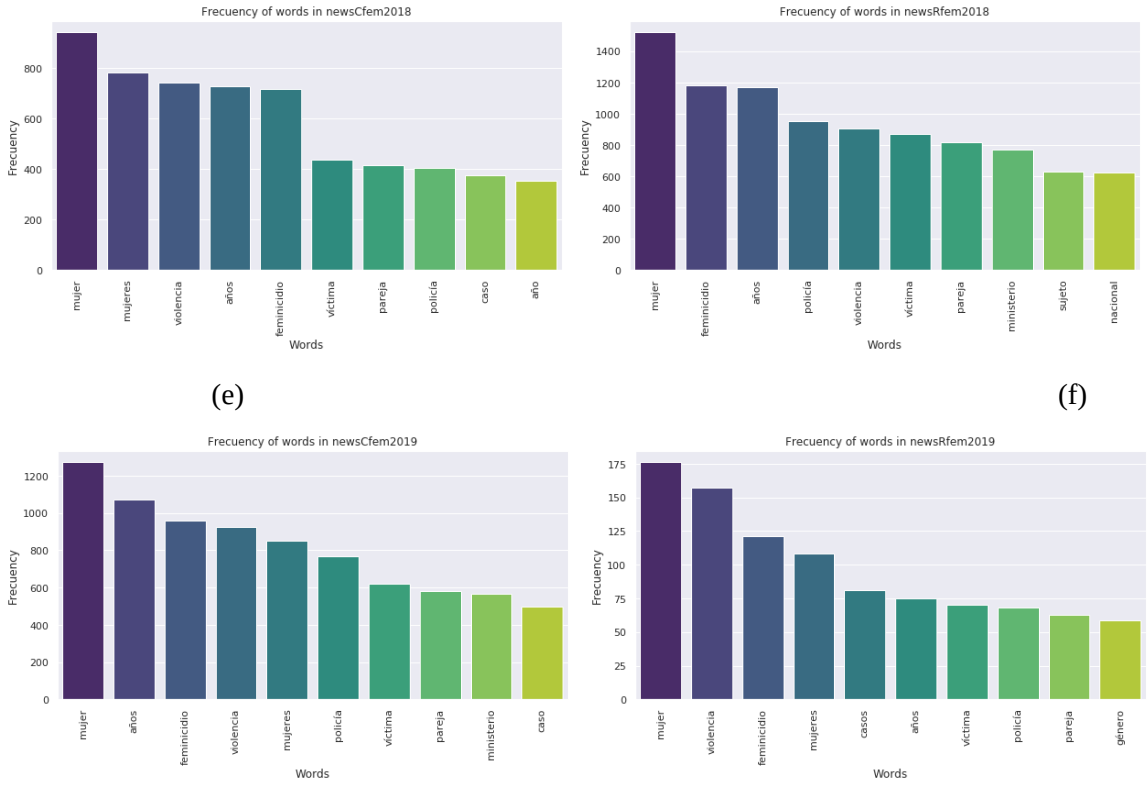


Fig. 4: Most frequent words in news of femicides and sexual abuse of 2 Peruvian newspapers: El Comercio (a) 2016, (c) 2017, (e) 2018, (g) 2019 and La República (b) 2016, (d) 2017, (f) 2018, (h) 2019

(a) (b)

Fig. 5: Spanish Verbs that appeared more frequently in news about (a) femicides (NewsCfem2019) and (b) sexual abuse (NewsCAbs2019).

(a) (b)

Fig. 6: Spanish Nouns that appeared more frequently in news about (a) femicides (NewsCfem2019) and (b) sexual abuse (NewsCAbs2019)

2) Unstructured Data Analysis: The case of the News dataset is different since the tools used in the preprocessing stage for unstructured data depend on the characteristics of the dataset and the model that is applied [6]. For the News data set, two keywords related to gender violence were found: "Sexual Abuse" and "Femicide" [34]. In this stage, we did not follow all the steps of the Natural Language Processing (NLP) used by the default models (tagger, a parser, and an entity recognizer), but we used some NLP techniques to make known the composition of each data set.

These data set do not have MVs, however, it is necessary to identify if that information is consistent and related to the topic under study, because when the scrapping process is performed, the keywords might not have a direct relation with the queries on the website.

As can be seen in Fig. 4, the most frequent and common words for the news that reports on cases of femicide and sexual abuse are shown.

To qualitatively show the frequency of appearance of verbs and nouns in news about femicides and sexual abuse, we used the same data sets for the newspaper El Comercio (NewsCfem2019, NewsCabs2019). The word clouds are shown in Figs. 5 and 6. The size of the word is directly related to the number of times it is mentioned. Lexical coherence is distinguished for the three types of search, and all the words are related to the context of studying gender inequality.

Once the three steps (Collecting, Parsing, and Pre-processing) are completed, the data can be used in the KDD process, through data mining techniques, allowing them to find patterns that show a deeper analysis.

IV. **RESULTS**

In this section, we show the characteristics and the content's details of each of those data sets.

Table 6 shows the data sets according to each of the family members considered by the Peruvian state in the ENDES survey. In contrast to the original data sets, they have the necessary columns and rows according to the exposed literature on gender studies in this work. Also, the cleaning process applied reduced the noise because "NaN" registers were managed. Regarding the indexation, original data sets needed many ID columns ("HHID", "CASEID2, "HHIDX", etc.) to join and associate the data between tables. After the pre-processing, it was possible to relate by index and other variables such as location. Further, this data could be easily used for any statistical software. All these results are visible comparing the 16 files for each year that we originally had. Finally, we obtained only 9, 4 for women, and the rest for the other members of the family, the description of the home, and the situation of violence.

Table 7 describes the clean information taken from the ENARES survey about women's characteristics in Peru and their perception of gender violence. The original datasets were found between different files and sections, they could hardly be related, their format could not be used in other statistical software than Stata or any that would accept 'DBF' type files. The process of cleaning the columns of unique data and MVs, allowed them to be used for increasing the possibilities of applying different models and tools on these data.

Table 8 refers to the payroll of workers in the Executive, Legislative and Judicial powers in Peru. These data are not publicly available in formats that do statistical analysis. Besides, they are dispersed by months of the year, and after the pre-processing, it is observed that they already belong to a single set of data per year in each state power. The anonymization of the data is not only interesting in this data set, but also imperative. A variable as important as gender would provide more information than people's names. However, the government prioritizes the transparency and publishes the names of public officials. On the other hand, people's gender could make these data sets a potential source of study on wage inequality.

Table 9 shows the records of complaints of domestic violence. This information was divided into data on the victim and the attacker, and according to the identifier of the complaint, it was possible to gather all the disperse data set into a single one. In the same way, the management of the MVs and the parsing of the format allow us to use these data in statistical software for deeper analysis.

Finally, Table 10 describes the news data sets in two important press outlets in Peru, related to sexual abuse and femicide cases. This information was not available in a viable format for its analysis. Then, the scraping process was essential, as well as the cleaning of the text and the identification of locations.

The need and importance of rigorous pre-processing carried out on all the data sets presented in this paper for the study of gender inequality is crucial to obtain accurate results when applying machine learning models. The new characteristics of all the analyzed and pre-processed data sets provided clean and consistent data in conditions and would be used in data mining techniques.

In the following section, we used two of the pre-processed data sets and applied linear models of machine learning. These models are sensitive to MVs since they assume that the data follow a normal distribution and as it was demonstrated in this research, MVs have a great impact in this regard [26]. One of the objectives when applying linear models is to obtain consistent results and acceptable validation metrics [51], [26], also demonstrating that the pre-process carried out was adequate and necessary.

IV. APPLICATIONS

A. Domestic violence prediction models in the interviewed households

Gender violence is any violent act or aggression within the framework of a system of domination relations of one gender over another (mainly female) and it is not only a structural cause, it is also a consequence of inequality. Therefore, promoting gender equality is a fundamental objective to reduce violence against women. Domestic violence against women is considered gender violence and, in most cases, it has to do with precarious situations of households [47].

In Peru, INEI conducts a voluntary survey to assess domestic violence through ENDES. Respondents who agree to testify if they are victims of violence, share in many cases, the same socio-economic characteristics as other women who refuse to do it. This behavior could alert to the presence of domestic violence in households that share similar characteristics and identify a percentage of the vulnerable population. As a result, public policies could be developed and government efforts could be directed towards focused attention.

In this work, the tool used and applied to the data collected and pre-processed for the study of gender inequality was classification algorithms, such as logistic regression [10]. The objective was to predict how many of the households that did not provide

information on domestic violence could be vulnerable to it. In this particular experiment, we used some data from the ENDES from the years 2016 to 2019. The data sets used were those of the "Home description" and the questionnaires about "Violence episodes" (Table 6). Starting from the "HHID" and "CASEID" identification cells, we constructed new data sets made up of those families that answered the violence survey and those that did not. When analyzing the antecedents recorded by the women who suffered some type of violence (physical, sexual, psychological, from the husband, or another relative), we decided to separate the level of aggression into low (class = 1) and high (class = 2), since the number of observations according to the type of violence varied for each respondent. Then, we split the dataset into the training set (70%) and the test set (30%) through a cross-validation algorithm. In Table 11 we show the implementation of the logistic regression model results, for the studied period. The accuracy of the algorithm reached an encouraging and reliable minimum value of 82%.

The potential of this linear model is that it can be extrapolated and would be applied to other data sets to be used in different scenarios.

TABLE VI
DATA SETS DESCRIPTION AND DETAILS WHICH BELONG TO THE ENDES SURVEY AFTER BEING PREPROCESSED

Family member	Dataset	Columns	Rows	Content	Type of variables
Women	FactIndMA16	10	70460	Age, Literacy, Ethnicity, Right Identity	Numerical Categorical
	FactIndMA17	10	70460		
	FactIndMA18	10	70460		
	FactIndMA19	10	70460		
	FactIndMB16	11	34996	Work, Nuptiality	Categorical
	FactIndMB17	11	34996		
	FactIndMB18	11	34996		
	FactIndMB19	11	34996		
	FactIndMC16	15	162903	Maternal mortality and Pregnancy	Categorical
	FactIndMC17	15	162903		
	FactIndMC18	15	162903		
	FactIndMC19	15	162903		
	FactIndMD16	86	38777	Contraception (use and methods)	Categorical
	FactIndMD17	86	38777		
	FactIndMD18	86	38777		
	FactIndMD19	86	38777		
Partner	FactPart16	5	11543	age, Literacy, Work	Numerical Categorical
	FactPart17	5	33163		
	FactPart18	5	34792		
	FactPart19	5	33111		
Children under 5 yesrs old	FactChildMin516	7	11543	Age, Sex, Right Identity, Weight, Height, Anemia	Numerical
	FactChildMin517	7	33163		
	FactChildMin518	7	34792		
	FactChildMin519	7	33111		

Children over 5 yesrs old	FactChildOv516	8	11543	Age, Sex, Literacy	Numerical
	FactChildOv517	8	33163		
	FactChildOv518	8	34792		
	FactChildOv519	8	33111		
Home description	Home16	72	11543	Socio-economical Houses 's characteristics	Categorical
	Home17	72	33163		
	Home18	72	34792		
	Home19	72	33111		
Violence episodes	Violence16	47	33136	Domestic violence episodes	Categorical
	Violence17	47	33169		
	Violence18	47	34792		
	Violence19	47	33878		

TABLE VII
DETAILS OF THE DATASETS IN ENARES SURVEY AFTER BEING PRE-PROCESSED

Family member	Dataset	Columns	Rows	Content	Type of variables
Women	WOCAP600	27	1599	Housing quality Sex, Age, Degree of instruction Members' marital status Sociodemographic characteristics of women 18 years of age or older married or in coexistence (women marital status) Physical, sexual and/or psychological violence of which women may be victims. Perceptions and attitudes towards gender roles	Categorical
	WOCAP100	31	1599		
	WOCAP200	30	6460		
	WOCAP3451	220	1599		
	WOCAP3452	220	1599		
	WOCAP3453	220	1599		
	WOCAP3454	220	1599		
	WOCAP3455	220	1599		
	WOCAP3456	220	1599		
	WOCAP3457	220	1599		
	WOCAP3458	220	1599		
	WOCAP3459	220	1599		
	WOCAP35510	220	1599		
	WOCAP35511	220	1599		
	WOCAP35512	220	1599		
	WOCAP35513	220	1599		
	WOCAP35514	220	1599		
	WOCAP35515	220	1599		
	WOCAP35516	220	1599		
	WOCAP35517	220	1599		
	WOCAP35518	220	1599		
	WOCAP35519	220	1599		
	WOCAP35520	220	1599		
	WOCAP35521	220	1599		
	WOCAP35522	167	1599		

TABLE VIII
DATASET DETAILS OF THE BRANCHES OF GOVERNMENT AUTHORITIES AFTER BEING PRE-PROCESSED

Batch Data	Data set	Columns	Rows	Content	Type of variables
Executive authority	Executive 2016	15	2985	Executive branch employees	Numerical and Categorical
	Executive 2017	15	2988		
	Executive 2018	15	2605		
	Executive 2019	15	2454		
Judicial authority	Judicial 2017	15	145455	Judicial branch employees	Numerical and Categorical
	Judicial 2018	15	126267		
	Judicial 2019	15	108987		
Legislative authority	Legislative 2016	15	37406	Legislative branch employees	Numerical and Categorical
	Legislative 2017	15	35929		
	Legislative 2018	15	18936		
	Legislative 2019	15	37381		

TABLE IX
DETAILS OF COMPLAINT DATASETS AFTER BEING PRE-PROCESSED

Batch Data	Data set	Columns	Rows	Content	Type of variables
Complaints dataset	SIDPOL_2016	51	133207	Complaints of domestic violence in Peru	Numerical and Categorical
	SIDPOL_2017	51	179610		
	SIDPOL_2018	51	179610		
	SIDPOL_2019	51	289569		

TABLE X
DETAILS OF NEWS DATASETS AFTER BEING PRE-PROCESSED

Batch	Data set	Columns	Rows	Content	Type of variables
newsRfem2016-19 newsCfem2016-19	newsRfem2016	6	750	News about femicide retrieved from web pages	Categorical (Text and dates)
	newsRfem2017	6	770		
	newsRfem2018	6	800		
	newsRfem2019	6	711		
	newsCfem2016	6	1215		
	newsCfem2017	6	998		

	newsCfem2018	6	1105		
	newsCfem2019	6	976		
newsRabs2016-19 newsCabs2016-19	newsRabs2016	6	1239	News about sexual abuse retrieved from web pages	Categorical (Text and dates)
	newsRabs2017	6	1500		
	newsRabs2018	6	1640		
	newsRabs2019	6	1003		
	newsCabs2016	6	934		
	newsCabs2017	6	1307		
	newsCabs2018	6	1435		
	newsCabs2019	6	1402		

TABLE XI
PERFORMANCE EVALUATION METRICS OF THE LOGISTIC REGRESSION MODEL APPLIED TO THE ENDES DATASET TO PREDICT DOMESTIC VIOLENCE ACCORDING TO SOCIOECONOMIC VARIABLES

Dataset	**Class**	**Precision**	**Recall**	**F1-score**	**Support**	**Accuracy**	**Size (test-sample)**
2016	1.0	0,82	0,84	0,83	2595	0,83	5206
2016	2.0	0,84	0,82	0,83	2611	0,83	
2017	1.0	0,83	0,85	0,84	2473	0,84	4981
2017	2.0	0,85	0,83	0,84	2508	0,84	
2018	1.0	0,86	0,82	0,84	2531	0,85	5123
2018	2.0	0,83	0,87	0,85	2592	0,85	
2019	1.0	0,85	0,79	0,82	2087	0,82	4185
2019	2.0	0,8	0,86	0,83	2098	0,82	

B. Experimentation on data sets related to gender inequality

The goal of machine learning is to create a tool that allows the solution of a task. It uses some of the variables or characteristics of an object to identify whether they belong to a certain group. In the case of the linear classifier, the objective is fulfilled by making a classification decision based on the value of a linear combination of these characteristics, which as they are vectored forming an array of characteristics. In this work, linear classification models were used through various algorithms. One of them was the classification of linear support vectors [10] which, in addition to generating a penalty if there is no success in the classification, has an optimal performance in samples with large amounts of data and also supports multiclass sets. Another of the classifiers used was the Ridge regression, which converts the target values to -1, 1, and then treats the problem as a regression task, with the same operation as a logistic regression [58]. The third algorithm used was the support vector machine with stochastic gradient descent (SGDClassifier) where the gradient function is estimated in each training sample updating the learning index incrementally [49]. Then, we applied the Perceptron algorithm, which like the SGDClassifier, is a binary linear classifier and shares the same [28] implementation form. Finally, we used the passive-aggressive classifier [7] whose implementation requires data that follows a normal distribution and which can then be represented in the form of a binary matrix.

The models mentioned are stable and have been used in implementations as part of the KDD process.

After applying the aforementioned classification algorithms, we obtained that in most of them, except for the Passive Aggressive Classifier in the data sets "2017" and "2018", the metrics such as the calculation of the F1-score were greater than 76% (Fig. 7). On the other hand, the LinearSVC classifier was the one that spent the most time training for the 2016 data set (0.4 seconds) (Fig. 8), but its testing time was one of the lowest for "2017" and "2018" (0.001 seconds) (Fig. 9). The slowest algorithm for the test time was the RidgeClassifier "2017" (more than 0.008 seconds), and the Perceptron algorithm (0.0005 seconds) was the fastest (Fig. 9).

Finally, using various classification algorithms, we obtained good metrics such as F1-score which was greater than 76% in most of them (except in Passive Aggressive Classifier in data set "2017"), Fig. 7.Other classification algorithms, such as the LinearSVC, lasted longer in training time (0.4 seconds) but its testing time was fast (0.001 seconds). The slowest algorithm for the testing task was RidgeClassifier (0.008 seconds) compared to the Perceptron algorithm (0.0005 seconds).

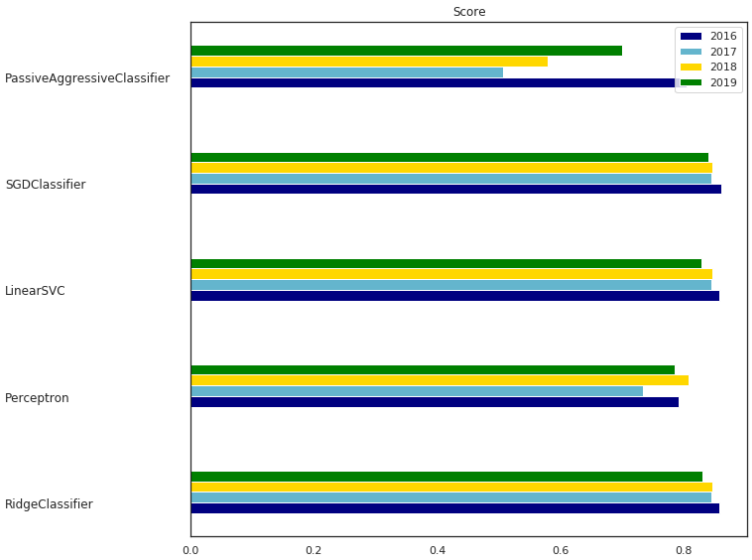


Fig. 7: Classification algorithms scores used for the study of gender inequality based on data set of Violence from the ENDES survey (2016-2019)

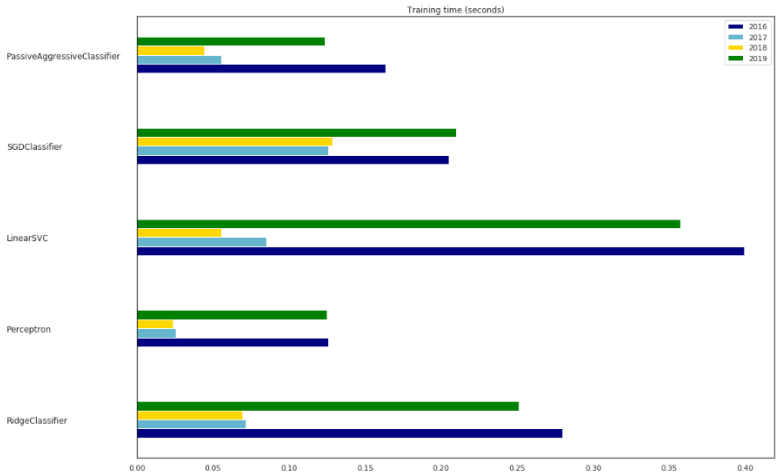


Fig. 8: Classification algorithms training time for the study of gender inequality based on data set of Violence from the ENDES survey (2016-2019)

It is important to note that the data sets "2017" and "2018" of the ENDES batch for the study of domestic violence prediction had very similar percentages of MVs, in addition to having a similar amount of data and their behaviors also follow the same trend. Data sets "2016" and "2019" had fewer MVs.

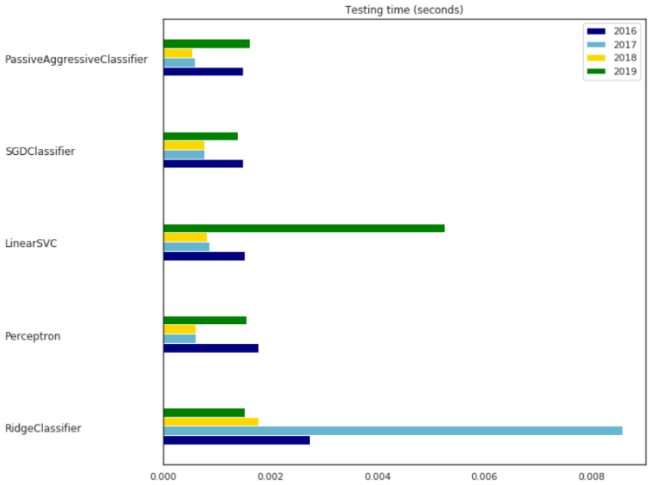


Fig. 9: Classification algorithms testing time for the study of gender inequality based on data set of Violence from the ENDES survey (2016-2019)

IV. CONCLUSIONS

Data pre-processing is a crucial step to avoid errors and extract valuable information. Without data pre-processing, these data errors would survive and lower the quality of data mining. Although the literature that addresses and describes data pre-processing techniques is vast; it presents scattered and general methods. In this current research, the role of pre-processing techniques in the study of gender inequality was evaluated. Then, we examined the performance of several known algorithms using pre-processing options on four data sets, finding optimal prediction metrics.

Finally, we would like to add that the incorporation of the pre-processing techniques carried out in this work for gender inequality analysis could enhance the visibility of this phenomenon that affects everyone through clean and consistent data; and contribute to achieving a more inclusive society, promoting equality and equity through data that guide policy decisions based on evidence. All the pre-processing techniques used for this topic could also be extrapolated to any analysis in the area of data mining.

REFERENCES

[1] Bachman, R. and Saltzman, L. E. (1995). Violence against women: Estimates from the redesigned survey. US Department of Justice, Office of Justice Programs, Bureau of Justice.

[2] Bell, J. H., von Sturmer, J. R., Needham, R., et al. (1969). The Elementary Structures of Kinship. Eyre and Spottiswoode.

[3] Bloss, J. (2017). The wiley blackwell encyclopedia of gender and sexuality studies. Reference Reviews.

[4] Bosco, C., Alegana, V., Bird, T., Pezzulo, C., Bengtsson, L., Sorichetta, A., Steele, J., Hornby, G., Ruktanonchai, C., Ruktanonchai, N., et al. (2017). Exploring the high-resolution mapping of gender-disaggregated development indicators. Journal of The Royal Society Interface,14(129):20160825.

[5] Bronfenbrenner, U. (1997). The ecology of cognitive development: Research models and fugitive findings. College student development and academic life: psychological, intellectual, social, and moral issues.

[6] Charniak, E. et al. (2016). Parsing as language modelling. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2331–2336.

[7] Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (2006). Online passive-aggressive algorithms. Journal of Machine Learning Research, 7(Mar):551–585.

[8] Dillon, S. and Collett, C. (2019). Ai and gender: Four proposals for future research. Technical report, University of Cambridge.

[9] Domínguez-Serrano, M. and Blancas, F. J. (2011). A gender wellbeing composite indicator: The best-worst global evaluation approach. Social Indicators Research, 102(3):477–496.

[10] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. Journal of machine learning research, 9(Aug):1871–1874.

[11] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11):27–34.

[12] Ford, D. A., Bachman, R., Friend, M., and Meloy, M. L. (2002). Controlling violence against women: A research perspective on the 1994 vawa's criminal justice impacts: (511962006-001). In Washington, DC: US Department of Justice, volume NIJ v.5.

[13] Fuentes, L. and Cookson, T. P. (2020). Counting gender (in) equality a feminist geographical critique of the 'gender data revolution'. Gender, Place & Culture, 27(6):881–902.

[14] Garcia, S., Luengo, J., and Herrera, F. (2015). Data Sets and Proper Statistical Analysis of Data Mining Techniques. In Data Preprocessing in Data Mining, pages 19–38. Springer.

[15] Garcia, S., Luengo, J., and Herrera, F. (2016a). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowledge-Based Systems, 98:1–29.

[16] Garcia, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., and Herrera, F. (2016b). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1):9.

[17] Gaye, A., Klugman, J., Kovacevic, M., Twigg, S., Zambrano, E., et al. (2010). Measuring key disparities in human development: The gender inequality index. Human development research paper, 46:41.

[18] González González, Á. and Alonso Cuervo, I. (2015). Manual práctico parala identificación de las desigualdades de género.

[19] Hajian, S. and Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. IEEE transactions on knowledge and data engineering, 25(7):1445–1459.

[20] Haroon, D. (2017). Statistics and probability. In Python Machine Learning Case Studies, pages 1–43. Springer.

[21] Haworth-Brockman, M. and Isfeld, H. (2009). Elementos para un análisis de género en las estadísticas de salud para la toma de decisiones. Organización Panamericana de la Salud.

[22] Heise, L. L. (1998). Violence against women: An integrated, ecological framework. Violence against women, 4(3):262–290.

[23] Hernando, A. (2017). The fantasy of individuality: on the sociohistorical construction of the modern subject. Springer.

[24] Huang, J., Gates, A. J., Sinatra, R., and Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. Proceedings of the National Academy of Sciences,117(9):4609–4616.

[25] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning, volume 112. Springer.

[26] Kelleher, J. D., Mac Namee, B., and D'arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press.

[27] Klugman, J. (2011). Human development report 2011. sustainability and equity: A better future for all. Sustainability and Equity: A Better Future for All (November 2, 2011). UNDP-HDRO Human Development Reports.

[28] Liu, W., Gao, P., Wang, Y., Yu, W., and Zhang, M. (2019). A unitary weights based one-iteration quantum perceptron algorithm for non-ideal training sets. IEEE Access, 7:36854–36865.

[29] Manning, C. D., Schütze, H., and Raghavan, P. (2008). Introduction to information retrieval. Cambridge university press.

[30] Metsis, V. and et al. (2006). Spam filtering with naive bayes – which naivebayes? In THIRD CONFERENCE ON EMAIL AND ANTI-SPAM (CEAS).

[31] Mignoli, G. P., Siboni, B. P., Rignanese, P., Valentini, C., Toschi, T. G., et al. (2018). Ugii–university gender inequality index. a proposal from the university of bologna. Technical report, Center for Open Science.

[32] MIMP-Perú (2016). Plan nacional contra la violencia de género. Technical report, Ministerio de la mujer y poblaciones vulnerables.

[33] Ochoa Rivero, S. (2002). Factores asociados a la presencia de violencia hacia la mujer. Lima. pages 15–16.

[34] Olivares, E. and Incháustegui, T. (2011). Modelo ecológico para una vida libre de violencia de género. Comisión Nacional para Prevenir y Erradicar la Violencia contra las Mujeres: México DF Recuperado de: http://www.conavim.gob.mx/work/models/CONAVIM/Resource/309/1/images/MoDecoFinalPDF.pdf.

[35] Pappalardo, L., Pedreschi, D., Smoreda, Z., and Giannotti, F. (2015). Using big data to study the link between human mobility and socio-economic development. In 2015 IEEE International Conference on Big Data (Big Data), pages 871–878. IEEE.

[36] Pateman, C. (1989). 'god hath ordained to man a helper': Hobbes, patriarchy and conjugal right. British Journal of Political Science, pages 445–463.

[37] Pedreschi, D., Ruggieri, S., and Turini, F. (2009). Integrating induction and deduction for finding evidence of discrimination. In Proceedings of

the12th International Conference on Artificial Intelligence and Law, pages157–166.

[38] Penner, A. M. (2015). Gender inequality in science. Science, 347(6219):234–235.

[39] Permanyer, I. (2010). The measurement of multidimensional gender inequality: Continuing the debate. Social Indicators Research, 95(2):181–198.

[40] Programme, U. N. D. (2018). Human development indices and indicators: 2018 statistical update. New York: UNDP.

[41] Publications, U. N. (2018). Human Development Indices and Indicators: 2018 Statistical Update. United Nations Development Programme.

[42] Romaní, F., César, G., and José, R.-C. (2011). Endes–autoexamen de mama en mujeres peruanas: prevalencia y factores sociodemográficos asociados. Análisis de la encuesta demográfica de salud familiar (endes). Technical report, Anales de la Facultad de Medicina Vol.72. No1. UNMSM. Facultad de Medicina.

[43] Rubin, G. (2012). Thinking Sex: Notes for a Radical Theory of the Politics of Sexuality. In Deviations: A Gayle Rubin Reader. Duke University Press.

[44] Ruggieri, S., Pedreschi, D., and Turini, F. (2010). Data mining for discrimination discovery. ACM Transactions on Knowledge Discovery from Data (TKDD), 4(2):1–40.

[45] Sammut, C. and Webb, G. I. (2017). Encyclopedia of machine learning and data mining. Springer.

[46] Sapounas, D. (2009). Heterogeneous data collection and data mining platform. US Patent App. 12/112,705.

[47] Segato, R. L. (2003). Las estructuras elementales de la violencia: contrato y status en la etiología de la violencia. Universidad de Brasilia, Departamento de Antropología.

[48] Sokoloff, N. J. and Dupont, I. (2005). Domestic violence at the intersections of race, class, and gender: Challenges and contributions to understanding violence against marginalized women in diverse communities. Violence against women, 11(1):38–64.

[49] Tang, Y. (2013). Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239.

[50] UNESCO (2020). Artificial intelligence and gender equality. Technical report, UNESCO.

[51] Urbanowicz, R. J. and Moore, J. H. (2015). Extracts 2.0: description and evaluation of a scalable learning classifier system. Evolutionary intelligence, 8(2):89–116.

[52] Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. Journal of Statistical Software, Articles, 45(3):1–67.

[53] Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In Proceedings of the 4th international conference onthe practical applications of knowledge discovery and data mining, pages 29–39. Springer-Verlag London, UK.

[54] Women, U. (2018). Turning promises into action, gender equality in the 2030 agenda for sustainable development. New York.

[55] Yang, H. (2018). Data preprocessing.

[56] Yin, X., Han, J., Yang, J., and Philip, S. Y. (2006). Crossmine: Efficient classification across multiple database relations. In Constraint-Based mining and inductive databases, pages 172–195. Springer.

[57] Zabokrtsky, Z. (2015). Feature engineering in machine learning. Institute of Formal and Applied Linguistics, Charles University in Prague.

[58] Zhang, L. and Suganthan, P. N. (2017). Benchmarking ensemble classifiers with novel co-trained kernel ridge regression and random vector functional link ensembles [research frontier]. IEEE Computational Intelligence Magazine, 12(4):61–72.

[59] Zhang, S., Zhang, C., and Yang, Q. (2003). Data preparation for data mining. Applied artificial intelligence, 17(5-6):375–381.

[60] Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery, 31(4):1060–1089.