Using Classification to Identify Potential Victims and Perpetrators of Spousal Violence in Pakistan

Abstract

This paper explores the application of Data Mining in identifying potential victims and perpetrators of spousal domestic violence among couples in Pakistan. Domestic violence deeply plagues the Pakistani society, given the country's general disposition towards gender inequality. To combat domestic violence, it is useful to identify couples most susceptible to it. Data of thousands of couples across various demographics within Pakistan was obtained from the Pakistan Demographic and Health Survey 2017-2018. Using the flags for various degrees and frequencies of domestic violence reported in the data, supervised learning was employed to predict whether a husband is subjecting/ will subject his wife to a certain type of domestic violence. The results of classification using J48 Decision Tree and a Naïve Bayesian Classifier were compared, for one of 17 selected output variables, i.e. identifiers of domestic violence. While the results were promising, it may be advisable to explore further nuances before applying the selected classifiers to the remaining output variables. Further investigation is needed to build a highly reliable and robust classification system.

Literature review

Pakistan is largely a patriarchal country, ranked at 151 of 153 countries in terms of gender gap index in 2020 [1]. Moreover, it has been ranked among the six most dangerous countries for women [2]. Domestic Violence is a highly concerning issue in the country. Around 34% of ever married women in Pakistan suffer from spousal violence. Many of these women incur injuries from violence, ranging from cuts and bruises to broken bones and miscarriages [3].

In order to improve the current situation, it is important to understand the factors that might be corelated and/ or causal to the phenomenon of domestic violence in Pakistan. Examining causality can help us to determine the reasons behind domestic violence in Pakistan and to work towards addressing these reasons. However, this is a highly nuanced task that would involve a deep dive into the sociology and psychology surrounding interactions within the Pakistani community. A suitable starting point would be to first study the patterns that are common within spousal violence cases in Pakistan, along with the profiles of perpetrators and victims. Moreover, using these patterns one can identify and target potential victims and perpetrators. This may help in prevent certain future violent episodes or save victims from potentially ongoing domestic violence.

This paper examines the use of data mining for identifying said patterns and constructing a classifier to predict whether spousal violence is likely among a couple. In the past, domestic violence records have been useful in identifying criminals. Holcomb and Sharpe [23] include domestic violence as a feature to forecast crimes. In recent years, Machine Learning has played a notable part in combating gender-based violence. Classification has been used with voice recognition systems and also real-time data from wearables used to identify anomalies in victims' speech and physical activity to predict situations of potential aggression [4,5]. While access to real time sensor data, as opposed to survey data, can essentially help overcome the biases due to factors like false reporting, consistently acquiring real time data may not be practical within all demographics of the Pakistani setting. Pakistan is ranked at 75 out of 79 countries in the global connectivity index [7] Access to technology may pose a challenge to several demographics within the country.

With regards to predicting Intimate Partner Violence and identifying perpetrators, Petering [8] uses SVM and Random Forest Trees on data from self-administered questionnaires over a longitudinal study, conducted with young homeless people in Los Angeles. Research in tapping into the vast potential of data as an answer to domestic violence is rather limited, despite the availability of data.

The Pakistan Demographic and Health Survey [3] acquired responses from over 3000 couples across a variety of demographics in Pakistan. This was the data source used in building our classifier. The Demographic and Health Survey has been an invaluable source of data for several localized applications of predictive analysis and has been used with Machine Learning algorithms to extract meaningful information. Neville [9] uses Causal Bayesian Networks on India's Demographic and Health survey to investigate the factors associated with childhood diarrhea. In a more relevant context, Amusa [10] uses decision trees on domestic violence data on the same survey in South Africa to predict vulnerability of women to Intimate partner violence.

Methodology

UNDERSTANDING THE DATA

A stata dataset was obtained after acquiring rights from the DHS (Demographic Health and Survey) program. The data was transferred to excel for initial observation and preprocessing. The survey for our dataset included participants from all provinces, rural and urban background, different age groups, wealth stratas, education and professional environments. The familial dynamics of the couples were noted with details such as, respondent's relationship with head of the family, number of children, polygamy dynamics (e.g. rank among wives), whether or not the wife has the autonomy to refuse sex/ demand that the husband uses condoms, who manages the finances, any jointly owned property etc. Additionally, the husband and wife were separately given hypothetical scenarios and asked if they consider it justified for the wife to be beaten in said scenario (e.g. if the wife argues with husband, if she burns the food, etc.). Moreover, it was inquired whether or not the husband demonstrates any controlling behaviors e.g. accusing the wife of infidelity, controlling who she meets, surveillance of where she goes. Lastly, the survey collected information of various kinds of emotional, physical and sexual abuse that the wife had previously experienced because of the husband and/ or others (e.g. ever been slapped, ever been pushed/ thrown something at by husband, ever been forced to perform sexual acts by husband, ever had miscarriages due to spousal violence during pregnancy, ever had broken bones, teeth or eye injuries due to spousal violence). These reports of abuse were all initially considered for potential outputs for the classifier.

PREPROCESSING

The following steps were employed:

a) Removing Records with Missing Attributes

Out of a total of 3334 tuples, only 2557 contained records for reports of abuse. These remaining tuples were separated from our data set.

b) Removing redundant Column

Columns that did not convey meaningful data were removed. These included: duplicate columns, multiple serial numbers, aggregated weights for which the calculations were not available hence could not be replicated. The data set contained other health and demographic attributes e.g. several columns detailing history with Malaria and Hepatitis. These were removed as we already had 100+ other potential attributes to pick and choose from which seemed more sensible. Of course, correlations with these attributes can be explored in a separate study, however, for the scope of the current study, the input attributes were limited to demographics and familial dynamics whereas potential output classes were the ones with reports of abuse.

c) Adjusting Data Types and Ranges

After the initial filtering there were a total of 117 attributes and 27 potential classes remained. Flags of emotional abuse such as "ever been insulted by husband" "ever been threatened by husband", etc were initially kept in both, the attribute (independent variable) list and the class list (dependent variable) because one can use the other attributes to predict likelihood of emotional abuse however, emotional abuse can also be used as a feature to identify physical abuse. Eventually, it was decided to keep emotional abuse markers as attributes for physical and sexual violence rather than being used as classes themselves. Attributes were now to be filtered on the basis of how their relationship with output classes. Since most of the variables were already categorical, the few remaining such as "Age" and "Number of children…" grouped into categorical variables. Missing cells were populated keeping in mind the context of each cell. For example, "Age of respondent at first birth" was initially blank for couples who didn't have children. These blanks were replaced with "Not Applicable". If a substantial number of records for one attribute had missing values with no discernible context, they were categorized as "Not Given". These records were separated from the rest during statistical analysis. Sparse missing values (e.g. 2-3 missing values in 2557 records) were populated manually with the most common label with similar features in other columns. In cases where the categories were poorly divided, (e.g. 2-3 records in one category vs thousands in another), the categories were redistributed (e.g. for "number of wives", 2 and 3 wives were merged into the same "more than one” category as there were very few records with 3 wives.

d) Statistical Analysis

To pick and choose the right variables, it was decided that a uniform statistical measure would be used for all variables. Since the variables were categorical and there was a large number of records, The Chi Square Test for categorical variables was used to obtain p-values. In this way, p-values were calculated for each of the 117 attributes vs each of the 27 potential output variables. (Please review the attached excel for preprocessing). 17 of the 27 class variables most relevant as indicators of domestic violence were kept while others were not processed further. The p-values of all attribute-class pairs were compared. Attributes with no/ very few significant classes and borderline significance (e.g. if attribute A has significant values for only 3/17 significant p-values, all between 0.05 and 0.01) were eliminated. Overlapping attributes were eliminated while retaining the more significant attribute that conveys the same information. In, the end, A list of 47 attributes were shortlisted. The Table 1 in Appendix A shows the shortlisted Attributes and Classes. For each Class, only the attributes marked with green are to be used for classification as these are the ones with significant p-values.

e) File formatting

The tool chosen for implementing classification was Weka. In order to process the data set, the excel files were converted to “.arff”. The data was loaded to Weka, visualized and classified (elaborated ahead). For the scope of the course project, only one Class Variable was classified. The rest of the outputs can be processed after further adjustments to current model and dataset.

VISUALIZATION

Class J1: HusbandPushedShookOrThrewSomething (at wife) was chosen as the dependent class variable. As it can be seen in Appendix A, Attribute I19: “decision maker for using contraceptives” was not significant so it was eliminated. The distribution of each attribute against this class was observed. The Figure 1 in Appendix B shows a selection of graphs each graph represents an attribute (e.g. age, etc) on the X axis, against which labels for the output class were stacked on the Y-axis. The output class had the following possible labels: “never”, “often”, “sometimes”, “yes-but-not-in-the-past-12 months”, shown in grey, red, teal and electric blue respectively. Appendix B shows a selection of the graphs which had the clearest visual indication that the distribution is not random. As the number of records that fall under each attribute are different, it would not make sense to compare the absolute heights of different colored sections of the bar. I.e. as a hypothetical example, if an attribute A has 577 records of class “yes” and 2000 records of class “no”. Suppose “yes” category has 100 records with label “sometimes” while “no” category has 150 records with said label. Naturally, the stacked bar for “no” would have a ticker red section than the bar for “yes”. However, when we translate these numbers into probabilities, they convey a different story. Given that a record falls in attribute A = “yes”, the probability of the husband from the record would push / shake / throw something at the wife from time to time is 100/577 which is higher than in the case where the record falls in A= “no” (probability p(J1=”sometimes” | A=no) is 150/2000). Hence, we compare the relative sizes of grey, teal, red and blue sections within each bar. So, if a larger **fraction of the bar** for A=yes is grey in comparison to fraction of bar for A=no, we can say that a husband falling in class A is more likely to “never” inflict this sort of domestic violence on his wife than a husband falling in class A=no. The metrics “violence never ratio”, “violence often ratio” and “violence sometimes” ratio capture this idea as given in appendix B. A higher “violence never ratio” means that there is a higher likelihood that a person in this attribute is more likely to “never” engage in violence of type J1. So high “violence never ratio”, lower “violence sometimes ratio” and lower “violence often ratio” are all indicators of a more peaceful class. The respective visual findings for each attribute are given in the figure below each attribute. The description isn’t being repeated here to avoid redundancy for the reader.

CLASSIFICATION

Classification was done on the Weka tool. 2 Classification models were compared. Firstly, as the raw data counts could be adequately used to compute conditional probabilities, Naïve Bayesian Classifier was employed. Secondly, as there was substantial evidence of decision trees being used for more accurate results the J48 Decision Tree was employed.

The outputs for various different splits were observed and 50-50 training and test data was chosen as the split with optimal accuracy. A higher percentage of training data is also less likely to overfit.

Decision tree is a discriminative model i.e. it models the decision boundary between classes whereas Naïve Bayesian is generative i.e. it explicitly models the actual distribution of a class. Decision trees are more flexible and can entertain data with missing values, outliers, and non-parametric feature interactions. However, they are more likely to overfit. Naïve Bayes is useful for applications where there are a large number of attribute. However, assumption of independence of random variables may result in loss inaccuracy. Our data has multiple attributes, no missing values but skewed distributions of output labels. Hence both algorithms stand a fair chance at performing, but might be likely to run into some issues because of the data distribution.

CLASSIFICATION INITIAL RESULT

Initially classification with Naïve Bayes yielded an accuracy of 75.86%. However, as seen in Table 2, the precision, recall and F-1 scores for "never" were very high (between 0.84 and 0.95) whereas those of the remaining attributes were very low (between 0.20 and 0.35). This is a clear sign of overfitting, as expected for a skewed data distribution.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
yes-but-not-in-last-12-months	0.321	0.049	0.221	0.321	0.262
Often	0.262	0.033	0.212	0.262	0.234
Sometimes	0.343	0.124	0.203	0.343	0.255
Never	0.841	0.310	0.935	0.841	0.885
Weighted Avg	0.758	0.275	0.820	0.758	0.785

Table 2: Naïve Bayesian results from initial classification

A similar observation was made for J-48 Decision Tree. The accuracy was better at 84.19% but the precision, recall and F-1 score showed an even worse case of overfitting.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
yes-but-not-in-last-12-months	0.132	0.024	0.194	0.132	0.157
Often	0.119	0.015	0.217	0.119	0.154
Sometimes	0.287	0.037	0.419	0.287	0.341
Never	0.961	0.552	0.902	0.961	0.931
Weighted Avg	0.842	0.469	0.809	0.842	0.823

Table 3: Decision Tree results from initial classification

The F-Measure for often and yes-but-not-in-the-last-12-minths is 0.157 and 0.154 respectively. As F-Measure is a harmonic mean that takes into account the uneven distribution of our data, it is a more indicative metric than accuracy. Moreover, our False Positive Rate for "never" is above 50% (0.552) while TP rate is above 95%. We can clearly see that our model is biased and is prone to wrongly categorize a great a lot of our data in the "never" category. I.e. a large proportion of actual perpetrators will not be identified and held accountable using this model. Hence the data had to be readjusted to account for this skew

POST PROCESSING AND ITERATING

After the initial classification, a second round of preprocessing was done. The skew of the imbalanced data was addressed by generating new instances of minority classes using SMOTE – Synthetic Minority Oversampling Technique. The quantity of each minority class (all except for "never") was increased by 100% by generating synthetic instances, using 5-nearest neighbors. Initially, the minority classes were increased by 100% but there was still a great degree of overfitting so then an increase of 50% was applied again. Once the new instances were added. The instances were then shuffled using the randomize filter in the Weka pre processing tab. Naïve Bayes and J48 Decision Tree were recreated, using a 50:50 split again.

Evaluation and Results

The Results for the Models were compiled again. The accuracy for Naïve Bayes dropped slightly to 74.83% and J48's accuracy dropped to 79.99%. However, the precision, recall and F measure scores drastically increased for both. F-measure for "often" went up to 0.770 for Naïve Bayes, while decision tree yielded an F-measure of 0.826 for the same. The lowest F-measures were 0.366 for Naïve Bayes and 0.496 for J48. False positive for “never” fell below 0.2 for both. Relative absolute mean for Naïve Bayes fell to 46% and J-48 to 41%. These values were touching 100% in the previous iteration. Overall, Decision tree had better, accuracy, precision, recall and F-measure than Naïve Bayes along with relative absolute mean. As stated earlier accuracy alone is not a good enough metric as our data is biased and accuracy does not reflect potential overfitting well hence a slight dip in accuracy is a small cost to pay for substantially better performance in all other mentioned domains. Table 5, 6 summarize the results described. The final Decision Tree can be seen in Appendix C. While the tree is rather cluttered, we can still see that the foremost decision was based on Emotional Violence attribute, ranking it as most significant, followed by Woman_ Afraid_of_Husband and Woman_Hurt_By_Others.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
yes-but-not-in-last-12-months	0.439	0.056	0.433	0.439	0.436
Often	0.832	0.061	0.717	0.832	0.770
Sometimes	0.496	0.106	0.427	0.496	0.459
Never	0.828	0.149	0.899	0.828	0.862
Weighted Avg	0.748	0.121	0.764	0.748	0.754

Table 5: Naïve Bayes Final Output

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
yes-but-not-in-last-12-months	0.645	0.048	0.565	0.645	0.157
Often	0.825	0.032	0.828	0.825	0.826
Sometimes	0.471	0.068	0.523	0.471	0.496
Never	0.899	0.182	0.887	0.889	0.888
Weighted Avg	0.800	0.131	0.799	0.800	0.799

Table 6: Decision Tree: Final Result

Limitation and Future work

While the results for second iteration are better, there is still substantial room for error while classifying. This means that perhaps a model better suited to imbalanced data should be employed. Moreover, data from previous DHS surveys may be helpful in producing better results. Upon finding the appropriate model, classification should be done for the remaining 16 attributes as well. The applications of this classifier extend beyond academia and can actually be extended into the real world to help people in escaping domestic abuse. Hence a viable future direction would be extending the research and developing and extending it to the Pakistan government sector for applications like policy making, targeting the right audience for protection programs and awareness programs. Furthermore, this data contains verbal reports from the users and are susceptible to misstatements. While determining p-values it was discovered that presence of husband or children during any section of the wife’s interview had a significant impact on the results. Hence effort needs to be made to acquire data from more reliable sources and to replicate the survey in a safer space for women, in which they can offer their candid input.

Conclusion

The data has successfully been used to gather insights on patterns in couples’ data related to domestic violence of type J1 and a classifier has successfully been made to identify the same. Adjusting the skew of the class variables has made the results less biased and helped us overcome overfitting. An effort needs to be made to find even better models to build an even more robust classifier. A similar classification should be then performed for classes J2 to J17

References

[1] Crotti, R., Geiger, T., Ratcheva, V., Zahidi, S. (2020). The Global Gender Gap Report, 2020. World Economic Forum, 2020

[2] Thomson Reuters Foundation. (2017). The world's most dangerous megacities for women. Retrieved from http://poll2017.trust.org/ [Google Scholar]

[3]National Institute of Population Studies and Macro International. (2018). *Pakistan demographic and health survey 2017–18*. Islamabad: National Institute of Population Studies and Macro International.

[4]Islam, A.; Akter, A.; Hossain, B.A. HomeGuard: A Smart System to Deal with the Emergency Response of Domestic Violence Victims. arXiv 2018, arXiv:1803.09401.

[5]. Hegde, N.; Bries, M.; Swibas, T.; Melanson, E.; Sazonov, E. Automatic recognition of activities of daily living utilizing insole-based and wrist-worn wearable sensors. IEEE J. Biomed. Health Inform. 2017, 22, 979–988. [CrossRef]

[6]Holcomb, J.P.; Sharpe, N.R. Forecasting police calls during peak times for the city of Cleveland. Case Stud. Bus. Ind. Gov. Stat. 2006, 1, 47–53.

[7]"See How the Digital Transformation of These Nations Compare." *Huawei*, www.huawei.com/minisite/gci/en/country-profile-pk.html#:~:text=Pakistan%20ranks%2076th%20in%20GCI,the%20global%20average%20this%20year.

[8]Petering, Robin & Um, Mee Young & Alipourfard, Nazanin & Tavabi, Nazgol & Kumari, Rajni & Nasihati Gilani, Setareh. (2018). Artificial intelligence to predict intimate partner violence perpetration.

[9]Neville Kenneth Kitson, Anthony C. Constantinou, Learning Bayesian networks from demographic and health survey data, Journal of Biomedical Informatics, Volume 113, 2021, 103588, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2020.103588.(https://www.sciencedirect.com/science/article/pii/S1532046420302161)

[10] Amusa, Lateef & Bengesai, Annah & Khan, Hafiz. (2020). Predicting the Vulnerability of Women to Intimate Partner Violence in South Africa: Evidence from Tree-based Machine Learning Techniques. Journal of Interpersonal Violence. 10.1177/0886260520960110.

Appendix A

Key	
yes	attribute Xi has been selected as input against the output YJ
no	attribute Xi has not been selected as input against the output YJ

	Input	Input Description	Selected	yes count		J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12	J13	J14	J15	J16	J17
					Outputs	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12	Y13	Y14	Y15	Y16	Y17	Y18	Y19	Y20	Y21
						ever been pushed, shook or had something thrown by husband/ partner	ever been slapped by husband/ partner	ever been punched with fist or hit by something harmful by husband/ partner	ever been kicked or dragged by husband/ partner	ever been strangled or burnt by husband/ partner	ever been threatened with knife/gun or other weapon by husband/ partner	ever been physically forced into unwanted sex by husband/ partner	ever been forced into other unwanted sexual acts by husband/ partner	ever had arm twisted or hair pulled by husband/ partner	ever been physically forced to perform sexual acts respondent didn't want to	experienced any less severe violence (d105a-c,j) by husband/ partner	experienced any severe violence (d105d-f) by husband/ partner	experienced any sexual violence (d105h-i,k) by husband/ partner	ever had bruises because of husband/ partner's actions	ever had eye injuries, sprains, dislocations or burns because of husband/ partner	ever had wounds, broken bones, broken teeth or other serious injury because of h	experienced any of listed husband/ partner's actions (d110x series)
I1	X1	Respondent Age Group	Y	10	sig	yes	yes	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	no	yes	yes
I2	X2	region	Y	14	sig	yes	yes	yes	yes	yes	yes	yes	no	yes	no	yes	yes	yes	yes	no	yes	yes
I3	X4	Husband Age Group	Y	8	sig	yes	yes	yes	no	no	no	no	no	yes	no	yes	no	no	yes	no	yes	yes
I4	X6	highest educational level	Y	13	sig	yes	yes	yes	yes	yes	no	no	no	yes	no	yes	yes	yes	yes	yes	yes	yes
I5	X8	relationship to household head	Y	8	sig	yes	yes	yes	yes	no	no	no	no	yes	no	yes	yes	no	no	no	no	yes
I6	X10	wealth index combined	Y	12	sig	yes	yes	yes	yes	yes	yes	no	no	yes	no	yes	yes	yes	yes	no	no	yes
I7	X11	educational level - husband	Y	11	sig	yes	yes	yes	yes	no	no	yes	no	yes	yes	yes	yes	yes	no	no	no	yes
I8	X12	total children ever born (group)	Y	10	sig	yes	yes	yes	yes	yes	no	no	no	yes	no	yes	yes	yes	no	no	no	yes
I9	X13	sons at home (group)	Y	12	sig	yes	yes	yes	yes	yes	yes	yes	no	yes	no	yes	yes	yes	no	no	no	yes
I10	X19	births in last 5 years (grouped)	Y	4	sig	yes	yes	no	no	no	no	no	no	no	no	yes	no	yes	no	no	no	no
I11	X21	age of respondent at 1st birth (group)	Y	10	sig	yes	yes	yes	yes	yes	no	no	yes	yes	no	yes	yes	no	no	no	no	yes
I12	X26	sons at home - husband (group)	Y	13	sig	yes	yes	yes	yes	yes	yes	yes	no	yes	no	yes	yes	yes	no	no	yes	yes
I13	X27	daughters at home - husband (group)	Y	9	sig	yes	yes	yes	yes	no	no	yes	no	yes	no	yes	yes	no	no	no	no	yes
I14	X35	number of women fathered children with	Y	7	sig	yes	yes	yes	no	no	no	no	no	yes	no	yes	yes	no	no	no	no	yes
I15	X42	respondent's rank among wives	Y	13	sig	yes	yes	yes	yes	yes	no	yes	yes	yes	yes	yes	yes	yes	no	no	no	yes
I16	X44	age at first sex (group)	Y	13	sig	yes	yes	yes	yes	yes	no	no	no	yes	no	yes	yes	no	yes	yes	yes	yes
I17	X49	age at first sex-husband (group)	Y	8	sig	yes	yes	yes	no	no	yes	no	no	yes	no	yes	no	no	no	yes	no	yes
I18	X53	husband's desire for children	Y	15	sig	yes	yes	yes	yes	yes	no	no	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
I19	X54	decision maker for using contraception	Y	10	sig	no	no	yes	yes	yes	yes	no	no	yes	no	yes	yes	no	yes	yes	no	yes
I20	X55	Not using contraception	Y	9	sig	yes	yes	yes	yes	yes	no	no	no	no	no	yes	yes	no	yes	no	no	yes
I21	X57	wife justified refusing sex: husband has other women	Y	13	sig	yes	yes	yes	yes	yes	yes	no	yes	yes	no	yes	yes	no	yes	no	no	yes
I22	X58	husband/partner's occupation (grouped)	Y	7	sig	yes	yes	no	no	yes	no	no	no	yes	no	yes	no	yes	no	no	no	yes
I23	X66	person who usually decides on visits to family or relatives	Y	11	sig	yes	yes	yes	yes	no	yes	no	yes	yes	yes	yes	yes	no	no	no	no	yes
I24	X67	person who usually decides what to do with money husband earns	Y	12	sig	yes	yes	yes	yes	yes	no	no	no	yes	no	yes	yes	yes	yes	no	yes	yes
I25	X68	beating justified if wife goes out without telling husband	Y	14	sig	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	yes	yes	yes	yes	no	no	yes
I26	X71	beating justified if wife refuses to have sex with husband	Y	15	sig	yes	yes	yes	yes	yes	no	yes	yes	yes	yes	yes	yes	yes	yes	no	yes	yes
I27	X72	beating justified if wife burns the food	Y	12	sig	yes	yes	yes	no	yes	no	yes	yes	yes	yes	yes	yes	yes	no	yes	no	no
I28	X73	beating justified for at least one specific reason	Y	14	sig	yes	yes	yes	yes	yes	no	yes	yes	yes	no	yes	yes	yes	yes	no	yes	yes
I29	X78	beating justified: wife goes out without telling husband - husband	Y	9	sig	yes	yes	yes	yes	no	no	no	no	yes	no	yes	yes	no	yes	no	no	yes
I30	X80	beating justified: wife argues with husband - husband	Y	9	sig	yes	yes	yes	yes	no	no	no	no	yes	no	yes	yes	no	yes	no	no	yes
I31	X84	owns a house alone or jointly-husband	Y	9	sig	yes	yes	yes	yes	yes	no	no	no	yes	no	yes	yes	no	no	no	no	yes
I32	X85	owns land alone or jointly-husband	Y	7	sig	yes	yes	no	yes	no	no	no	no	yes	no	yes	yes	no	no	no	no	yes
I33	X87	respondent can refuse sex	Y	10	sig	yes	yes	yes	yes	no	no	yes	no	yes	no	yes	no	yes	yes	no	no	yes
I34	X88	respondent can ask partner to use a condom	Y	11	sig	yes	yes	yes	yes	yes	no	yes	no	yes	no	yes	yes	no	yes	no	no	yes
I35	X89	presence of children < 10 for 'wife beating justified' questions	Y	11	sig	yes	yes	yes	yes	yes	yes	no	no	yes	no	yes	yes	no	yes	no	no	yes
I36	X90	presence of husband for 'wife beating justified' questions	Y	11	sig	yes	yes	yes	yes	yes	yes	no	no	yes	no	yes	yes	no	yes	no	no	yes
I37	X98	number of control issues answered 'yes' (d101x = 1)	Y	17	sig	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
I38	X99	respondent ever physically hurt husband/partner when he was not hurting her	Y	14	sig	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	no	no	yes
I39	X101	frequency of husband/partner being drunk	Y	15	sig	yes	yes	yes	yes	yes	yes	yes	no	yes	no	yes	yes	yes	yes	yes	yes	yes
I40	X102	no person other than husband/partner ever physically hurt respondent	Y	16	sig	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	yes	yes	yes
I41	X103	respondent's father ever beat her mother	Y	14	sig	yes	yes	yes	yes	yes	yes	yes	no	yes	no	yes	yes	yes	yes	no	yes	yes
I42	X104	respondent afraid of husband/partner most of the time, sometimes or never	Y	17	sig	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
I43	X106	type of relationship with husband	Y	9	sig	yes	yes	no	yes	yes	no	no	yes	yes	no	yes	yes	no	no	no	no	yes
I44	X107	had a say in choosing the husband	Y	8	sig	yes	yes	yes	no	no	no	no	no	yes	no	yes	no	no	yes	no	no	yes
I45	X112	justifies domestic violence: neglected in-laws	Y	15	sig	yes	yes	yes	yes	yes	no	yes	yes	yes	yes	yes	yes	yes	yes	no	yes	yes
I46	X113	beating justification: neglected in-laws - husband	Y	10	sig	yes	yes	yes	yes	yes	no	no	no	yes	no	yes	yes	no	yes	no	no	yes
I47	X117	experienced any emotional violence (d103x series)	Y	17		yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
			47			46	46	43	39	34	18	21	16	45	12	47	40	26	27	11	17	45

47 attributes chosen out of 117

Table 1: Selected Attributes and Classes

Appendix B

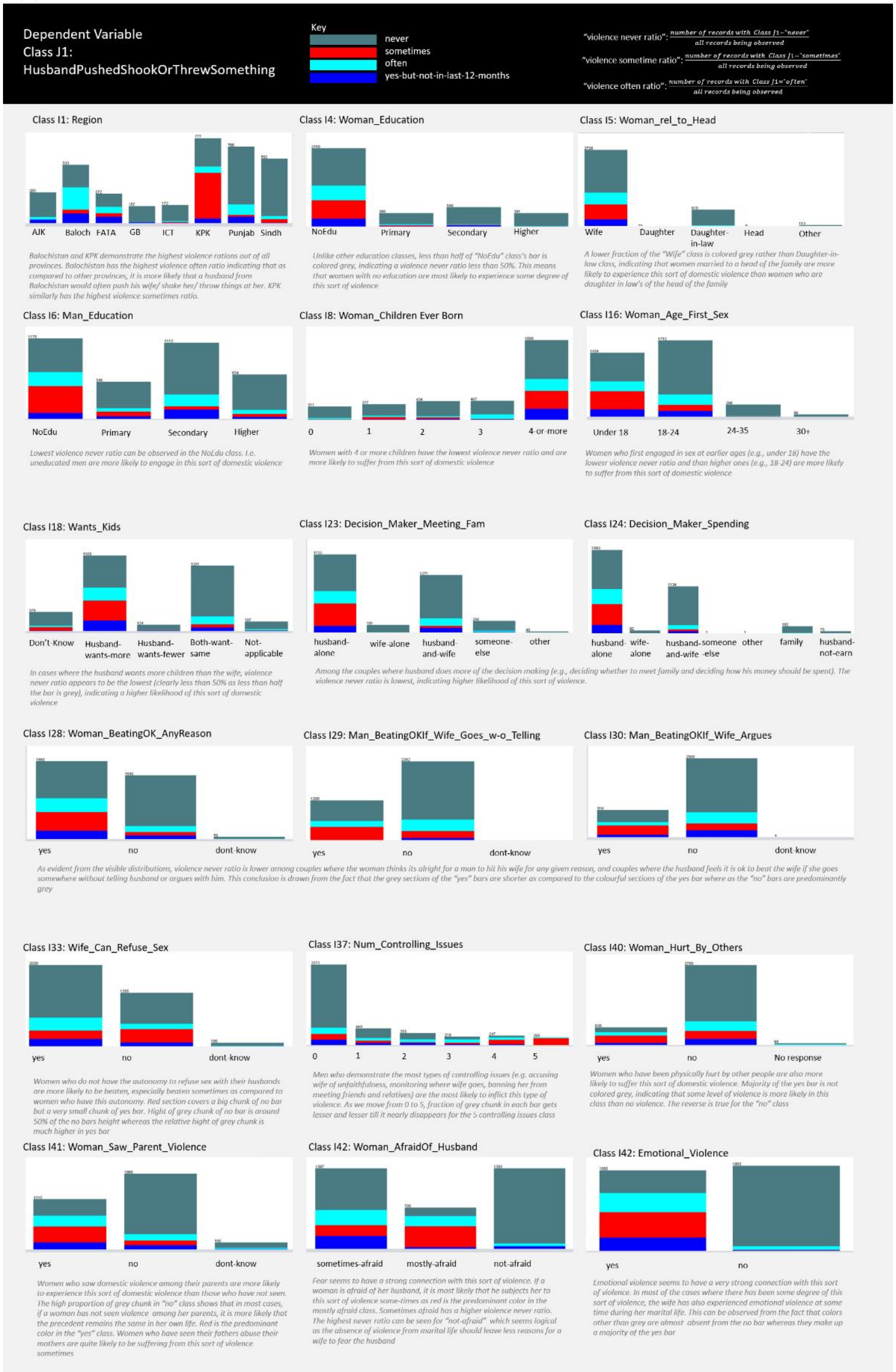

Figure 1: Visualizations of Attributes against Class J1

Appendix C

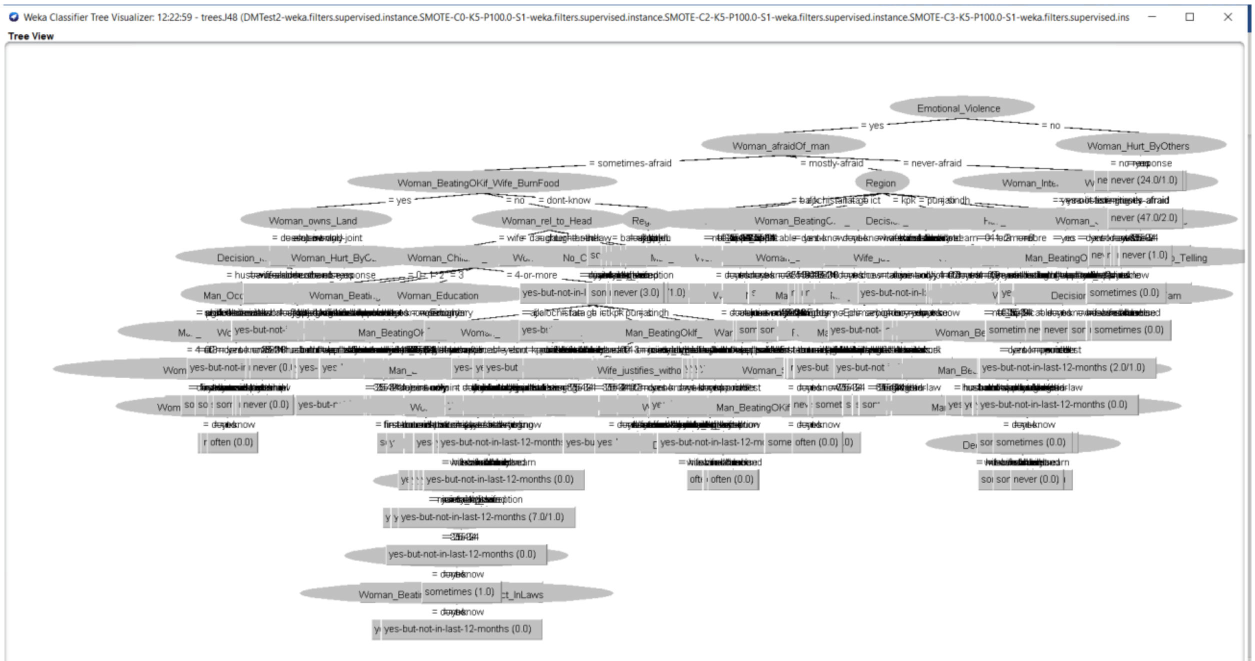

Figure 2: Final Decision Tree