Decision tree Model: *Predicting sexual offenders on the basis of minor and major victims*

Bhajneet Kaur
Research Scholar, AIIT
Amity University
Noida, UP
bhajneetahuja@gmail.com

Dr. Laxmi Ahuja
Professor, AIIT
Amity University
Noida, UP
lahuja@amity.edu

Dr. Vinay Kumar
Ex Scientist & Professor
Vice Chairman (Cum Chairman Elec)
CSI Delhi Chapter
vinay5861@gmail.com

Abstract*— Sexual offences spoil the whole culture of any society, city, state or country. So, the identification and the prediction of the sexual offenders are very important. This research paper proposed a model to predict the sexual offenders on the basis of major or minor victims which could help to take any kind of decision by police departments, sexual harassment cells, and law enforcement agencies to differentiate the sexual offenders of major or minor victims to enhance the implementation of security accordingly for crime prevention. The model first classifies the sexual offenders then does their prediction through the predictor variables age, race, and weight. To deploy the decision model overall dataset has been divided into 70:30 (training data: test data) ratio. The proposed decision tree model has been resulted with 79.8% accuracy rate with the 70% of test data and model validated through 30% of remaining data as given 79.1% accuracy rate. Model can predict 82.7% sexual offenders on the basis of minor victims and 75.5% on the basis of major victims.

Keywords*— sexual offenders, minor victim, major victim, Decision tree, Index values, Gain chart, response chart, machine learning, CHAID, SPSS.

I. Introduction

Sexual violence is the major issue for public health. It violates the mental, social, physical and reproductive health of a person whether it is against women or children. According to the World Health organization report 30% of the women faced sexual or physical violence by their intimate partners. Sexual offenders are those, who commit or do the sexual or physical crime. There are various researches and studies have been done on the sexual offenders but it is difficult to identify who offend against minor victim or major. Police departments and the various public and private law organizations are taking many initiatives to classify or recognize sexual offenders on the basis of major and minor victims but the issue has not been resolved even the sexual offences are reproducing day by day. There is a need of a system which can easily identify the offenders involved in such kind of offences so that the decision could be taken to punish the offenders. The decision tree, a classification technique has been used to develop the model for predicting the sexual offenders of minor and major victims on the basis of their age, race, height and weight. Decision tree is a predictive technique under the supervised leaning mechanism of machine learning. With the help of this technique the classification of sexual offenders has been done on the basis of the minor or major victim using the target variable victim_minor from the dataset. The overall paper structure has been divided into five segments. Introduction has been written in the Segment I. Segment II explains the review of literature related to sexual offenders or sexual crime and the proposed model technique. Segment III gives the description of dataset which used in this paper, proposed model summary and model interpretation. Performance measure of predictive model is analyzed in segment IV. Conclusion and future work is explained in segment V.

II. Related Study

Various studies have been conducted on sexual offenders and offences. [2] Proposed a study about the perception of female student in India on sexual behavior of coaches and sexual harassment, as resulted 31% of the female student reported unacceptable and serious occurrence in the sexual behavior of coaches. 55 studies conducted in 24 countries by [11], concluded there are 9 girls out of 100 are victims of sexual offences. Workplace sexual harassment is continuously facing by some women. On the basis of this a review has been published by [7] with the objectives of accumulate the awareness of sexual offences, to evaluate them and to identify the areas in which there is the need of investigation. A study has been conducted in Canada on sexual offenders for female children to predict the behavior in responses to sexual stimuli, including pictures of nude females of various ages with the description of both consenting and non-consenting sex [14]. A study conducted by [18] discussed that the offender who offends older women and the offender who offends children are of different age groups. There are various types of crime against women i.e. sexual offences, kidnapping, murder etc. Various factors are identified by [21] affecting crime against women, also checked the impact of the factors on crime rate.

III. Proposed Work

Dataset Specification: The dataset used in the decision tree model is sourced from the portal of Chicago police department as secondary data. Chicago police maintained the list of sexual offender's data who residing in the Chicago city. Dataset has been extracted from Citizen Law Enforcement Analysis and reporting system developed by the police department of the Chicago city. Dataset updates on daily basis and it has been categorized into various attributes. After the compilation of data some attributes are

978-1-5386-9346-9/19/$31.00 ©2019 IEEE

taken for the implementation of the decision tree model. The target variable of this dataset includes as victim minor or victim major because the sexual offenders are categorized into two values Y (1) and N (0). Y means the sexual offender is of minor victim and N means the sexual offender is of major victim. In dataset the attribute name of the target category is Victim_minor. Other attributes used in the decision tree model as independent variables are Race, Age, Height and weight of the sexual offenders. There are various methods to implement decision tree model but in the proposed model CHAID (chi-squared Automatic Interaction Detection) method has been used as a growing method of the tree because it is very effective method for the large dataset and uses multi-way splits.

Model Specifications: After the compilation of the dataset four independent variables have been taken for the proposed model i.e. Race, Age, Height & Weight and attribute named Victim_Minor chosen as a dependent variable. The overall dataset has been categorized into the ratio 70:30 (training : test). 70% of the overall data used to develop the model and 30% data to test the model by using the split sample validation technique of CHAID method. As shown in the Table-1, maximum depth of the tree is 3 and minimum cases in the parent and child nodes are 100 and 50.

Model Results Interpretation: As per the result of the model summery of decision model mentioned in Table I, 3 predictors are found, out of 4 independent variables i.e. Race, Age and weight. Height could not found as a good predictor according to the current dataset. So, height has been excluded from the model. In the decision tree total 8 nodes have been formed. In 8 nodes, 5 nodes are terminal nodes and the depth of the tree depicted as 2.

The tree diagram is generated in the Figure 1 for the training data and in Figure 2 for the test data. As shown in Figure 1, the tree model has been developed for the training data (70%) by using CHAID method. Race of the sexual offenders is found as the best predictor (Node 0) of entire model out of 4 independent variables, which is classified into two nodes i.e. Node 1 & Node 2. Races like White Hispanic, White Asian /pacific islande, black Hispanic comes within Node 1 and only Black race counted in the Node 2. Next best predictor found as Age. Node 1 has been classified into 3 another nodes i.e. Node 3, Node 4 & Node 5. Sexual offenders predicts as in Node 3, those Age<=39, Age greater than 39 & less than 48 found in the Node 4 and Age>48 are counted in Node 5. Nodes: 3, 4, 5 are the terminal nodes i.e. there is no further classification of these nodes.

TABLE I. DECISION TREE MODEL SUMMARY

Specifications	Growing Method	CHAID
	Dependent Variable	victim_minor
	Independent Variables	RACE, AGE, HEIGHT, WEIGHT
	Validation	Split Sample
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	RACE, AGE, WEIGHT
	Number of Nodes	8
	Number of Terminal Nodes	5
	Depth	2

For Age <=39, 90.8% are sex offenders of minor victim and 9.2 % are of major victim. Age between 39 to 48 the minor victims are 62% and major are 37%. For the Race, White Hispanic, white Asian /pacific islande, black Hispanic and the Age > 48, 86.4% sexual offenders are of victim_minor. Since there is no child node below the Age variable so that these are considered as terminal nodes.

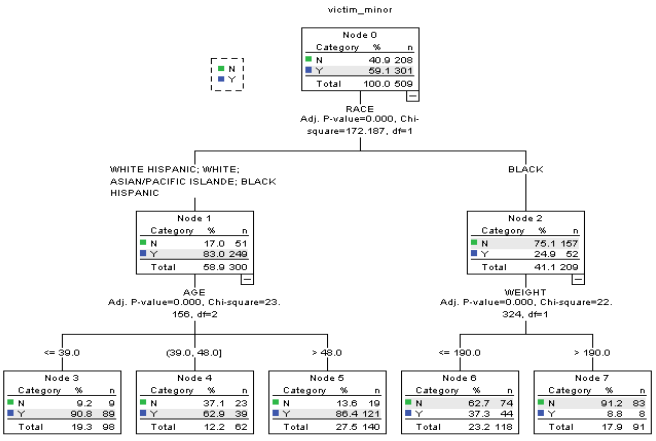

FIGURE 1. DECISION TREE MODEL FOR TRAINING DATASET

When Race equals to black i.e. Node: 2, then weight variable has been predicted the next predictor. When Race is Black and weight <=190 then 37.3% of the sexual offenders are of minor victim and 62.7 % are of the major victim's offenders. But when Race=Black and weight>190 then only 8.8% of the sexual offenders found under the victim –minor category and 91.2 found under the victim-major category.

As shown in the figure 2, the tree model has been formed with the test data i.e. 30% dataset of the entire data. In figure 2, classification of the independent variables has been done same as figure 1. With the Race: White Hispanic, white Asian /pacific islande, black Hispanic, Age has been categorized into 3 parts i.e. Node-3,4,5 with the following rules.

Race: White Hispanic, white Asian /pacific islande, black Hispanic and Age<=39 then Vctim_minor=70.7%.

Race: White Hispanic, white Asian /pacific islande, black Hispanic and 39<Age<=48 then Vctim_minor=88.5%.

Race: White Hispanic, white Asian /pacific islande, black Hispanic and Age>48, then Vctim_minor=85.7% and victim_major 14.3%

Race: Black and Weight < 190, then Vctim_minor=32.7% and victim_major 67.3%.

Race: Black and Weight > 190, then Vctim_minor=15.2% and victim_major 84.8%.

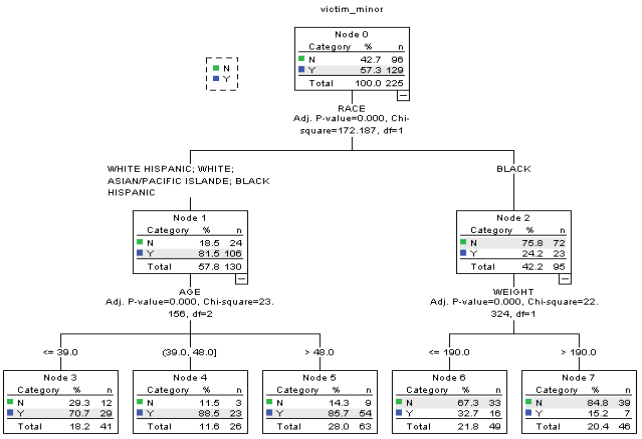

FIGURE 2. DECISION TREE MODEL FOR TEST DATASET

There are total 7 nodes formed for the training and test dataset. For every node the significant value has been found ass .000 that means which is less than .0001 for all splits in the model for both the data. In the predicted category both the values are given for the different nodes i.e. victim minor or major. For training and test dataset variables are given in the table with their chi-square value at every node and degree of freedom. In the tree table every node has found with the significance value so, the proposed tree model is significantly valid.

IV. PERFORMANCE MEASURE

The performance of the decision tree model is measured by the Gain chart and Index chart. Table II depicts all the information regarding terminal nodes of Training data and Test data as per the mentioned columns node, gain, response and index. Node N is value of total number of cases occurred in both the categories of target variable. In figure 1, terminal Node 7 presented value N=91as total number of cases participated in this particular node so that for Node 7, 91 value mentioned in column N in table 3 and the percent indicates total percentage covered by N with the entire cases i.e. for N=91, percent=17.9%. Gain is defined by the total number of cases in each terminal node in the target category. During the development of the decision tree model the target category dependent variable was chosen as 'N' means "non-minor victim" i.e. "major victim". So, Gain column is divided into two parts: 'N' & 'Percent'. If N =83 for node-7, then 83 cases could be correctly predicted by the decision model for major victim when race=black and weight>190.0 because the target category is defined as major victim. Gain percent is the overall percentage with the gain cases (N). Response is defined by the total percentage of the target category of the decision tree model as shown in the Figure 1 for the particular node. As shown in the Table II, 91.2% response has been depicted for Node-7 as per the training data because 91.2% sexual offenders are classified on the basis of major victim when race of the sexual offender = black and weight > 190.0. Index has been shown as last column of Table II which measured by the ratio of response percentage of the target category compared to the entire sample. If index > 100%, then there are more cases in the target category than the overall percentage in the target category.

TABLE II. TERMINAL NODE TABLE FOR CALCULATING GAIN, RESPONSE AND INDEX PERCENTAGE

Sample	*Node*	*Node*		*Gain*		*Response*	*Index*
		N	Percent	N	Percent		
Training	7	91	17.9%	83	39.9%	91.2%	223.2%
	6	118	23.2%	74	35.6%	62.7%	153.5%
	4	62	12.2%	23	11.1%	37.1%	90.8%
	5	140	27.5%	19	9.1%	13.6%	33.2%
	3	98	19.3%	9	4.3%	9.2%	22.5%
Test	7	46	20.4%	39	40.6%	84.8%	198.7%
	6	49	21.8%	33	34.4%	67.3%	157.8%
	4	26	11.6%	3	3.1%	11.5%	27.0%
	5	63	28.0%	9	9.4%	14.3%	33.5%
	3	41	18.2%	12	12.5%	29.3%	68.6%

Gain chart has been shown in Figure 3, produced by the gain percentage from Table II. For the training data & test data, gain charts are almost equal so it validates the overall results. Response chart has also been depicted in Figure-4 with the responses of training data and test data. Response of the Node-7 has shown as 91.2% and for Node-3 it has been shown as 9.2%. So, the response chart has been formed in declined way from 91.2% to 9.2%. Same as the index chart has been depicted in Figure 5 for training and test data.

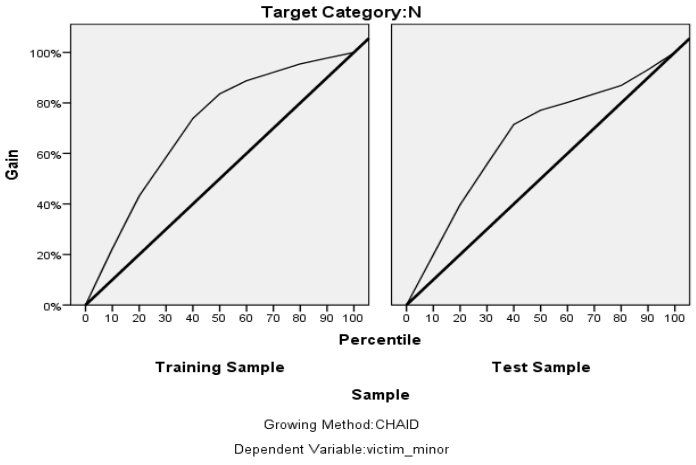

Figure 3. Gain chart for Training and Test Data

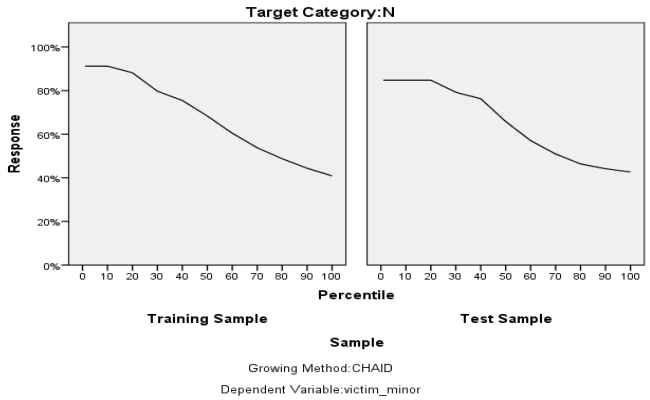

Figure 4. Response Charts for Training and Test Data

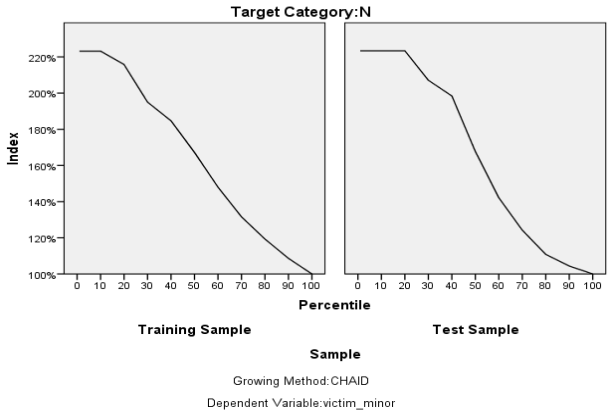

Figure 5. Index charts for Training and Test Data

TABLE III. RISK ESTIMATION FOR TRAINING AND TEST DATA

Sample	*Risk Estimate*	*Std. Error*
Training	.202	.018
Test	.209	.027

It has been found through the Table III, 20.2% risk could be there for using this model to predict the sexual offenders. Risk has also been validated by the test data in Table III because model developing risk and model validating risk are approximating same. So, there could be approx. 20% chances of wrong prediction but accuracy chances are 79.8% as measured as mentioned in Table IV through the 70% training data. Accuracy rate to predict sexual offenders of non-minor i.e. majors is 75.5% and 82.7% found to predict the accurate results of minors, It means 24.5% offenders of non-minor i.e. major category are inaccurately classified by the model and 17.3% offenders of minor category are inaccurately classified by the model. With the remaining 70% of test data, model has been validated by using split validation method, validation resulted 79.1% overall correctness, includes 75% offenders of major and 82.2% offenders of minors.

TABLE IV. MODEL'S PREDICTED ACCURACY RESULTS FOR TRAINING AND VALIDATION

Classification				
Samples	**Observed**	**Predicted**		
		N	N	Y	Percent Correct
Training	N	157	51	75.5%
	Y	52	249	82.7%
	Overall Percentage	41.1%	58.9%	79.8%
Test	N	72	24	75.0%
	Y	23	106	82.2%
	Overall Percentage	41.1%	58.9%	79.8%

V. CONCLUSION AND FUTURE SCOPE

This paper has been proposed a decision model to predict the sexual offenders on the basis of minor and major victims. The model is analyzed for its accuracy using Gain chart and Index chart and it is found that decision tree model gives almost good results for the classification of sexual offenders. 70% of the entire data has been used to develop the model, called training data set which resulted 79.8% of accuracy rate and it has also been validated through remaining 30% of the dataset. Model validated successfully because through the test data overall 79.1% rate of accuracy has been delivered which is very close to the accuracy rate of training data. For the development of the decision tree model in-total 4 variables i.e. Age, weight, Height, Race were involved initially but only 3 variables are considered by the technique as good predictors i.e. Race, Age, weight for the development of decision tree. There is 20.2% of risk to use this model but accuracy result is 79.8%. This model can be used by the police departments, various law organizations to identify and predict the sexual offenders based on major and minor victims and implement the security accordingly.

REFERENCES

[1] Burgess, A. W., Commons, M. L., Safarik, M. E., Looper, R. R., & Ross, S. N. (2007). Sex offenders of the elderly: Classification by motive, typology, and predictors of severity of crime. Aggression and Violent Behavior, 12(5), 582-597.

[2] Ahmed, M. D., van Niekerk, R. L., Ho, W. K. Y., Morris, T., Baker, T., Ali Khan, B., & Tetso, A. (2018). Female student athletes' perceptions of acceptability and the occurrence of sexual-related behaviour by their coaches in India. International Journal of Comparative and Applied Criminal Justice, 42(1), 33-53.

[3] Burgess, A. W., & Clements, P. T. (2006). Information processing of sexual abuse in elders. Journal of Forensic Nursing, 2(3), 113-120

[4] Burgess, A. W., Dowdell, E. B., & Brown, K. (2000). The elderly rape victim: Stereotypes, perpetrators, and implications for practice. Journal of Emergency nursing, 26(5), 516-518.

[5] Collins, P. G., & O'Connor, A. (2000). Rape and sexual assault of the elderly–an exploratory study of 10 cases referred to the Irish Forensic Psychiatry Service. Irish Journal of Psychological Medicine, 17(4), 128-131.

[6] Sehgal, R., Mehrotra, D., & Bala, M. (2018). A Decision Tree Approach to Identify the Factors Affecting Reliability for Component-Based System. In Smart Computing and Informatics (pp. 237-243). Springer, Singapore.

[7] McDonald, P. (2012). Workplace sexual harassment 30 years on: A review of the literature. International Journal of Management Reviews, 14(1), 1-17.

[8] Yang, H., & Fong, S. (2018). Incremental Optimization Mechanism for Constructing a Balanced Very Fast Decision Tree for Big Data. In Innovative Research Methodologies in Management (pp. 111-144). Palgrave Macmillan, Cham.

[9] Jaafari, A., Zenner, E. K., & Pham, B. T. (2018). Wildfire spatial pattern analysis in the Zagros Mountains, Iran: A comparative study of decision tree based classifiers. Ecological Informatics, 43, 200-211.

[10] Quist, J., Mirza, H., Cheang, M. C. U., Telli, M. L., Lord, C. J., Tutt, A. N. J., & Grigoriadis, A. (2017). Association of a four-gene decision tree signature with response to platinum-based chemotherapy in patients with triple negative breast cancer.

[11] Barth, J., Bermetz, L., Heim, E., Trelle, S., & Tonia, T. (2013). The current prevalence of child sexual abuse worldwide: a systematic review and meta-analysis. International journal of public health, 58(3), 469-483.

[12] Shechory Bitton, M., & Ben Shaul, D. (2013). Perceptions and attitudes to sexual harassment: an examination of sex differences and the sex composition of the harasser–target dyad. Journal of Applied Social Psychology, 43(10), 2136-2145.

[13] Kim, J. W., Lee, B. H., Shaw, M. J., Chang, H. L., & Nelson, M. (2001). Application of decision-tree induction techniques to personalized advertisements on internet storefronts. International Journal of Electronic Commerce, 5(3), 45-62.

[14] Marshall, W. L., Barbaree, H. E., & Christophe, D. (1986). Sexual offenders against female children: Sexual preferences for age of victims and type of behaviour. Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 18(4), 424.

[15] Wu, D. (2009). Supplier selection: A hybrid model using DEA, decision tree and neural network. Expert Systems with Applications, 36(5), 9105-9112.

[16] Terry, K. J. (2015). Sex offender laws in the United States: smart policy or disproportionate sanctions?. International Journal of Comparative and Applied Criminal Justice, 39(2), 113-127.

[17] Bettinger-Lopez, C. (2018). Violence against Women Normative Developments in the Inter-American Human Rights System.

[18] Browne, K. D., Hines, M., & Tully, R. J. (2018). The differences between sex offenders who victimise older women and sex offenders who offend against children. Aging & mental health, 22(1), 11-18.

[19] Bhaskaran, S. S., Lu, K., & Aali, M. A. (2017). Student performance and time-to-degree analysis by the study of course-taking patterns using J48 decision tree algorithm. International Journal of Modelling in Operations Management, 6(3), 194-213.

[20] Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy, 32(9), 1761-1768.

[21] Kaur, B., Ahuja, L., & Kumar, V. (2018). Factors Affecting Crime Against Women Using Regression and K-Means Clustering Techniques. In Industry Interactive Innovations in Science, Engineering and Technology (pp. 149-162). Springer, Singapore.

[22] Giguere, R., & Bumby, K. (2007). Female sex offenders. Policy and Practice Brief]. Center for Sex Offender Management, USA Retrieved from http://www. csom. org/pubs/female sex offenders brief. pdf.

[23] Gavankar, S., & Sawarkar, S. (2015, December). Decision Tree: Review of Techniques for Missing Values at Training, Testing and Compatibility. In Artificial Intelligence, Modelling and Simulation (AIMS), 2015 3rd International Conference on(pp. 122-126). IEEE

[24] Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), 660-674.

[25] Bettinger-Lopez, C. (2018). Violence against Women Normative Developments in the Inter-American Human Rights System.

[26] Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy, 32(9), 1761-1768.

[27] Brackenridge, C. (1997). HE OWNED ME BASICALLY...' Women's Experience of Sexual Abuse in Sport. International Review for the Sociology of Sport, 32(2), 115-130.

[28] Mesarić, J., & Šebalj, D. (2016). Decision trees for predicting the academic success of students. Croatian Operational Research Review, 7(2), 367-388.

[29] Bachman, R., & Saltzman, L. E. (1994). Violence against women (Vol. 81). Washington, DC: US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics