Predicting sexual offenders using exhaustive CHAID techniques on victim's age

Bhajneet Kaur*

Amity University,
Noida (RDIAS), Uttar Pradesh, India
Email: bhajneetahuja@gmail.com
*Corresponding author

Laxmi Ahuja

Amity Institute of Information Technology (AIIT),
Amity University, Noida,
Uttar Pradesh, India
Email: lahuja@amity.edu

Vinay Kumar

Computer Society of India (CSI) Delhi Chapter,
Delhi, India
Email: vinay5861@gmail.com

Abstract: Sexual offences can spoil the whole culture of the society. This research paper proposes two decision models to classify and predict the sexual offenders of minor and major victims on the basis of their physical attributes namely age, race, weight and height using CHAID and Exhaustive CHAID techniques of decision tree. Overall dataset has been divided into 70:30 for building and testing the models. As resulted 79.8% rate of accuracy found by model using CHAID technique even model tested with 79.1% rate of accuracy. By using Exhaustive CHAID, 79.9% rate of accuracy depicts by the model developed through 70% of test data and model validated through 30% of test data with 78.8% rate of accuracy. The proposed models can help to take any kind of decision further by police departments, sexual harassment cells and law enforcement agencies for security purposes.

Keywords: sexual offenders; minor victim; major victim; decision tree; index values; gain chart; response chart; machine learning; CHAID; SPSS; exhaustive CHAID.

Reference to this paper should be made as follows: Kaur, B., Ahuja, L. and Kumar, V. (2022) 'Predicting sexual offenders using exhaustive CHAID techniques on victim's age', *Int. J. Applied Management Science*, Vol. 14, No. 1, pp.71–89.

Biographical notes: Bhajneet Kaur is Research Scholar at Amity Institute of Information Technology, Amity University, Noida. She is also working as an Assistant Professor in Rukmini Devi Institute of Advanced Studies, New Delhi affiliated to GGSIPU. She has teaching and research experience in the field of Computer Science, Information Technology and Management. She has

presented research papers in various National and International conferences, also received the best paper award. Her research work got published in National and International journals including Scopus enlisted journal. She has also been invited as a resource person in research field in various faculty development programs.

Laxmi Ahuja (PhD, CSE) is working as Dy. Director in Amity Institute of Information Technology for past 20 years. Her areas of interest are search engine, data mining and soft computing approaches. She has published more than 80 research papers in international and national conferences and journals. She successfully filed number of patents under IT domain by Patent Department, Govt. of India. She organised various sessions in international conferences and session chair in various conferences. She has supervised three students for PhD. She is Member of IEEE and Life Member of Computer Society, India, Past Vice chairman of IETE - Senior member of IACSIT.

Vinay Kumar did his PhD from University of Delhi, MCA from Jawaharlal Nehru University, Delhi. He worked as scientist in National Informatics Centre (NIC), Government of India for approximately 21 years. He also worked with Vivekananda Institute of Professional Studies (VIPS), affiliated to GGSIPU, as Professor and Dean of the IT Department and Research and Publication. He has extensively contributed to the knowledge arena through his books and research publications. He has published over 100 research papers in refereed journals and authored a book of *Discrete Mathematics* and a memoire *Killed Instinct*.

This paper is a revised and expanded version of a paper entitled 'Decision tree model: predicting sexual offenders on the basis of minor and major victims' presented at the 'Amity International Conference on Artificial Intelligence (AICAI)', Dubai, United Arab Emirates, 4–6 February 2019.

1 Introduction

Sexual offence is one of the major issues in the society. It destroys the social, mental and physical state of a person, whether it is done against women, men or children. As stated in the report published by world health organisation (WHO, 2014) 30% of women experienced sexual violence through their intimate partners. Sexual offenders are those criminals who commit physical offences. Various studies have been conducted on the basis of crime or criminal related issues that have not been resolved. It is difficult to distinguish and identify the offender on the basis of victim's age (minor or major). Various law organisations, whether private and public are implementing various security measures to protect the society from these vicious acts. But these issues are still irresolvable. These criminal activities are heard almost every day. Sexual offences are destroying society's mental health. There is an immediate need of a strong system against sexual offenders, so that they can be identified easily and punished accordingly. For classification of sexual offenders (of minor and major victims), a decision tree technique has been used to propose the model. The proposed model has been created on the basis of few independent variables like race, age, height and weight. Decision tree is one of the techniques that classifies and predicts which comes under the supervised leaning technique of machine learning. By using this technique various classification and predictive models have been developed. A predictive model has been developed by the

team of researchers in the medical science for the analysis of breast cancer (Quist et al., 2017). A model has been developed by (Upadhyaya et al., 2018) to classify the fake and genuine banknotes with very good rate of accuracy. Another work done by the group of the researchers (Oh et al., 2018) using decision tree to build a model to predict the risk of suicide attempt through intoxication in South Korea. As per the authors Intoxication is a very common suicidal method and the common reason of emergencies in medical centre in South Koria. In the proposed work, classification and prediction of sexual offenders-based upon minor or major victim, has been mentioned. The target variable defines the two different age groups of victims i.e. major or minor. There are various methods to build a model using decision tree technique. In the proposed models CHAID (chi-squared Automatic Interaction Detection) and Exhaustive CHAID methods have been used as growing methods of the decision tree because they are very effective and uses multi-way splits. Originally, CHAID algorithm had been proposed by Kass (1980) and Exhaustive CHAID by Biggs et al. (1991). Both CHAID and exhaustive CHAID consist of 3 basic steps to develop a tree – (i) merging, (ii) splitting & (iii) stopping. The tree grows by using these three steps in repetitive manner which starts from the root node. Two initial steps, splitting and stopping are same in both algorithms i.e. CHAID and Exhaustive CHAID. But third step namely "merging" uses an exhaustive search process to merge any similar kind of pair in iterative manner, until a single pair remains. Exhaustive CHAID is the modified version of the CHAID algorithm performs merging and testing of predictor variable thoroughly (Kass, 1980; Biggs et al., 1991; Goodman, 1979).

The formation of the entire paper is segregated into five components. Section 1 describes the introductory part. In Section 2, review of literature has been discussed related to current work i.e. sexual offenders or sexual crime and model technique. Model has been proposed in detail in Section 3 using CHAID and Exhaustive CHAID. Performance of the models has been measured in Section 4 on the basis of gain, response and index chart. Also, accuracy rate and risk estimation rate have been compared for each model. Conclusion has been drawn at last in Section 5 with future work.

2 Related study

Various research works have been done in the field of crime against women related to sexual offenders and offences. One of the studies had conducted to find the perception of female students against sexual harassment and offensive behaviour of coaches (Ahmed et al., 2018). As resulted, 31% of the female students reported unacceptability and serious occurrences in the sexual behaviour of coaches. Barth et al. (2013) discussed, 55 studies conducted in 24 countries which shows that out of 100, 9 girls are victims of sexual offences or crimes. Various studies have been directed upon the sexual harassment at workplace. McDonald (2012) had published his work by defining its objectives to acquire the awareness level of sexual offences among citizens, their evaluation and identification of those areas, where investigation is being needed. Marshall et al. (1986) conducted a study in Canada upon female children to predict the behaviour of sexual offenders against responses to sexual encouragements, via pictures or images of various age groups with the detailed information of both kinds of sex i.e. consenting and non-consenting. A related work has been done by Browne et al. (2018) that defines the age groups of offenders who offend older women and children. The crimes against women have been categorised into various kinds like sexual offences, murder, kidnapping, etc.

So there are various factors which are impacting crime or violence against women identified by Kaur et al. (2018). Also an idea has been proposed by Kaur et al. (2020) for developing a conceptual model using interpretive structural modelling (ISM) technique through the various factors.

3 Proposed work

Data set specification: For the analysis of sexual offenders a secondary dataset has been used, extracted from online portal of Chicago police department. The data set has been arranged in such a manner that it contains information of each sexual offender on the basis of certain variables. The source is an online portal known as "Citizen Law Enforcement Analysis and reporting system" developed and maintained by Chicago police on daily basis by keeping the track of updated records. Dataset has been distributed into various attributes like age, height, weight, race, etc. To deploy the classification and predictive model the dataset has been compiled first and after the compilation some of the required attributes are taken for the analysis purpose. Data set is specified with dependent and independent variables. The dependent or target variable of this data set is defined in two categories i.e. minor and major and on the other hand the sexual offender are categorised into two values Y (1) and N (0). Y means sexual assault upon a minor N means the same upon a major. As per the dataset target variable is named by Victim_minor. Other attributes are used as independent variables i.e. Race, Age, Height and weight of the sexual offenders to deploy the decision tree model.

Model specifications: Four independent variables are taken to propose the decision tree model i.e. Race, Age, Height and Weight and the dependent variable (target variable) has been named "victim_minor". To fulfil the assumption of the decision tree technique, the target variable must be of non-metric scale or categorical data. IBM SPSS 21 tool has been used for the analysis. For building the decision model extracted data set has been divided into 70:30 ratios as per training and testing. 70% of data set used to build the model and 30% data is used to test or validate the model. As shown by Table 1 – Model Summary Specialisation, maximum tree depth is 3 and minimum cases that came under the category of parent and child nodes are 100 and 50. The split validation method has been used to develop the model. Two kinds of methods have been used on dataset CHAID and Exhaustive CHAID. Exhaustive CHAID is the up-gradation of CHAID technique. Table 1 has shown the description about both the techniques on the basis of growth Method, dependent variables, independent variables, tree validation method, maximum tree depth and minimum cases occurred in parent and child nodes. For both type of growth methods i.e. CHAID and Exhaustive CHAID the maximum tree depth has been set as 3. Minimum number of cases that has been specified in the parent node is 100, whereas in the child node there are 50. Dependent variables have two categories Y and N. Y victim is a minor and N victim is a major.

Model results interpretation with CHAID and exhaustive CHAID: As resulted by model summary mentioned in Table 2 predictors (Race, Age and weight) are found, out of 4 independent variables in case of CHAID algorithm. Height has not been found as a good predictor by the CHAID method of decision tree technique. So, height variable has been excluded while development of the model. In the decision tree total 8 nodes have been formed. Out of 8 nodes, 5 nodes are generated as terminal nodes. Depth of the tree has been predicted as 2 by CHAID. The model summary of Exhaustive CHAID model

has also been given in Table 2. In this case only 2 variables are selected as best predictors instead of 3 i.e. Age and Race, out of 4 independent variables (Race, Age, Height and Weight). Here, decision tree model has been developed by using 5 nodes. 3 nodes have been drawn as terminal nodes. The depth of the decision model has been found as 2.

Table 1 Model summary specifications using CHAID and exhaustive CHAID

Specialisation	
Growth method	CHAID and exhaustive 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

Table 2 Model summary output using CHAID and exhaustive CHAID

Results		
Growth method	CHAID	Exhaustive CHAID
Independent variables included	Race, Age, Weight	Race, Age
Number of nodes	8	5
Number of terminal nodes	5	3
Depth	2	2

The other two variables i.e. Height and Weight have not been found as good predictors. So these two independent variables have been dropped while developing decision tree model using Exhaustive CHAID. Table 2 has found the model summary output of both the growth methods. Lesser number of independent variables has been involved in the formation of decision tree model of Exhaustive CHAID as compared to CHAID algorithm. As depicted, 5 terminal nodes have been formed in case of CHAID but only 3 are formed in case of Exhaustive CHAID. As resulted the depth of the tree is 2 in both the growth methods.

The tree diagram of decision model has been generated in Figure 1 by using 70% of training data as shown in Figure 2 for 30% of test data by CHAID algorithm. Figures 3 and 4 represent the tree diagrams of training and test data by using Exhaustive CHAID method of decision tree. The tree diagram representations are different for both CHAID and Exhaustive CHAID but rate of accuracy are almost same. As shown in Figure 1, Race variable has been found as the best predictor placed at root node i.e. Node 0. Further Race has been classified into two another nodes i.e. Nodes 1 and 2. Races of sexual offender like White Hispanic, White Asian /pacific island, black Hispanic are found at Node 1 and only Black race has been occurred at Node 2. Next best predictor found is "Age". Node 1 has been classified into 3 another nodes i.e. Node 3, Node 4 and Node 5. Sexual offenders predicts at Node 3, whose Age <=39, Age greater than 39 and less than 48 found at Node 4 and Age>48 are counted at Node 5. Nodes: 3, 4, 5 are the terminal nodes i.e. there is no further classification of these nodes.

Figure 1 Decision tree model for training data set using CHAID

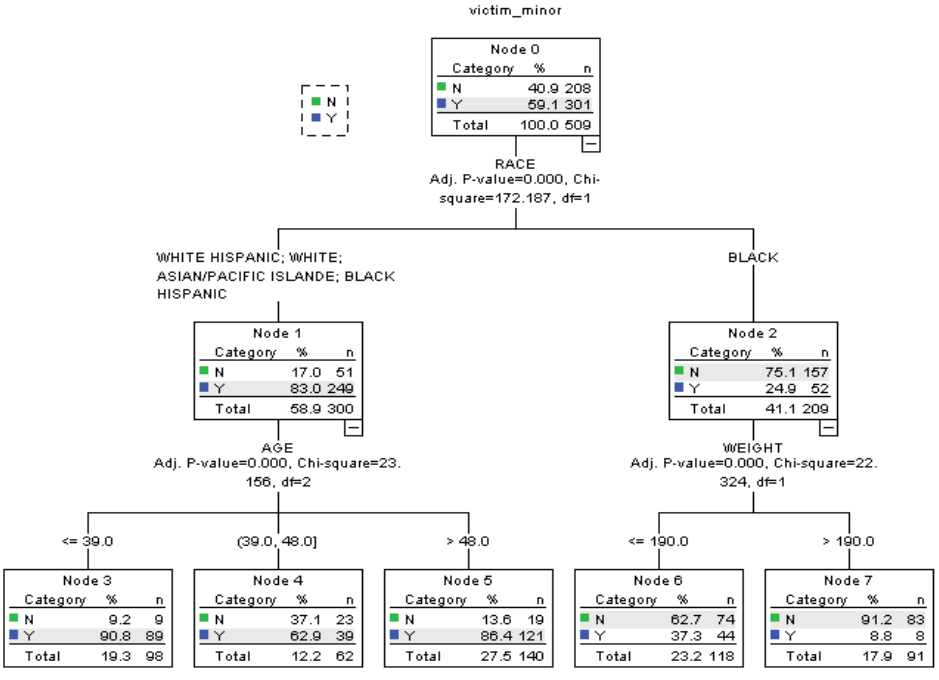
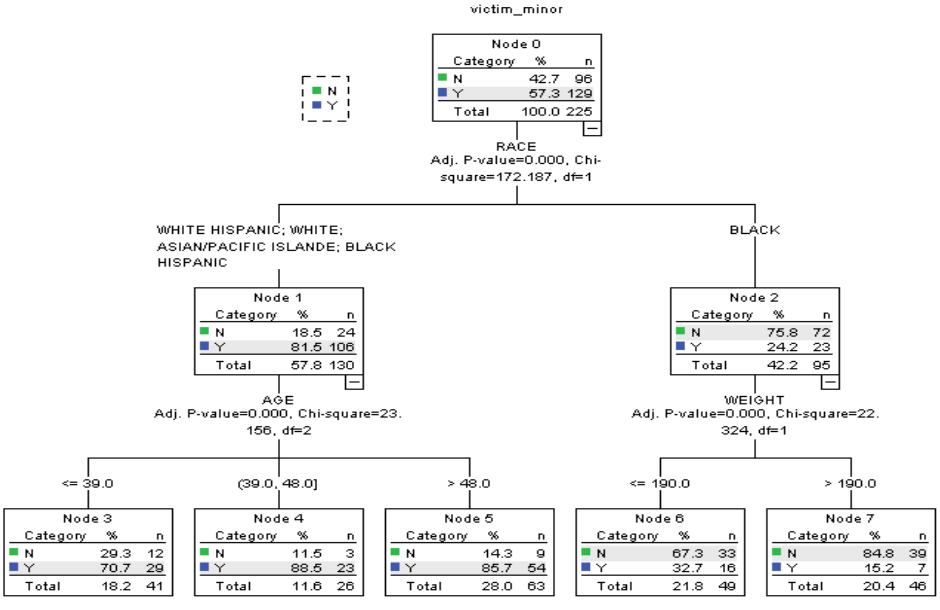
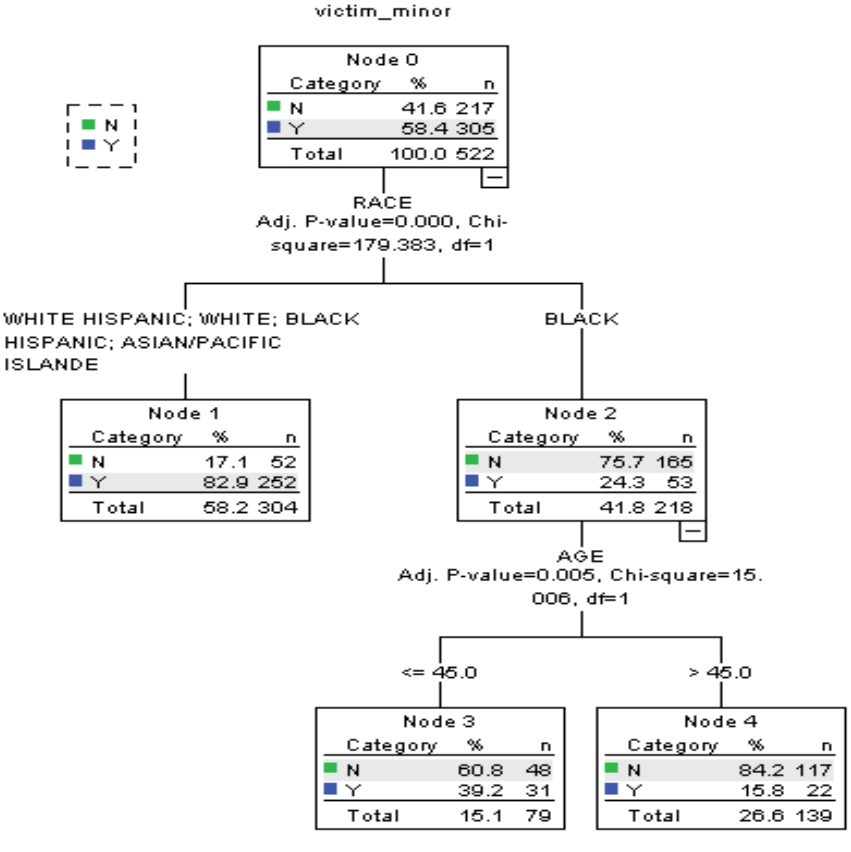


Figure 2 Decision tree model for test data set using CHAID



For Age <=39, 90.8% sexual offenders can be predicted for minor victim and 9.2% for the major victim. When Age is between 39 and 48, the minor victims can be 62% and major can 37%. As shown by Node 5 in Figure 1, when Race of the sexual offender is White Hispanic or white Asian /pacific island or black Hispanic and Age > 48, 86.4% then 86.6 % of offenders found under minor category and 13.6% are under major category. Since there is no child node below the Age variable that is why Nodes 3, 4 and 5 are considered as terminal nodes.

Figure 3 Decision tree model for training data set using exhaustive CHAID



Node 2 has been formed in Figure 1, when Race= Black, which is further divided into two more Nodes 6 and 7. Weight variable has been predicted as next predictor. When Race is Black and weight <=190 then 37.3% of the sexual offenders are found under the minor victim and 62.7 % are of the major victim. But when Race=Black and weight>190 then only 8.8% of the sexual offenders found under the minor victim category and 91.2% found under the major victim category.

Figure 4 Decision tree model for test data set using exhaustive CHAID

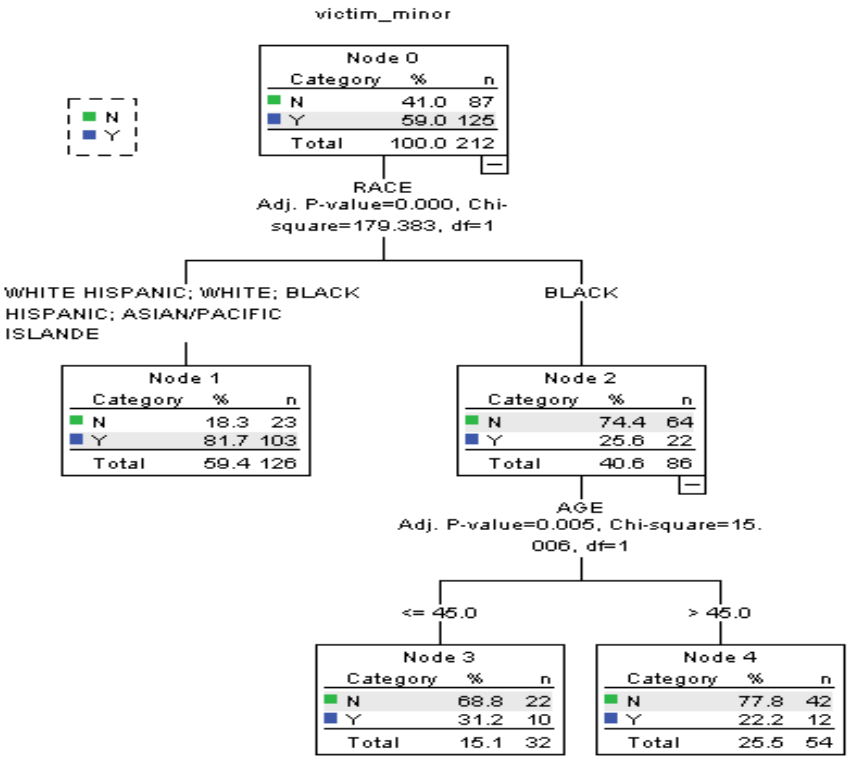


Figure 2 has been described the validated tree model developed through rest of the 30% of data set using CHAID growth method. In total 225 records of the entire data have been found, which has been divided as 96 major and 129 minor victims. Following decision rules can be formed to classify and predict the sexual offenders on the basis of the major and minor victims from Figure 1 and Table 3. The rate of prediction of each category has been given with generated decision rules.

The detailed information of every node of decision tree graph has been depicted in Table 3 via various columns as Node numbers, with total number of cases classified at every node for major category. Next column has been shown about the total number of cases found in the second category of the target variable: victim_minor. Then, percentage column depicts the percentage of minor category occurred from the overall number of cases. Further predicted category column has been given. In the next column the parent node number has been depicted for every node. For Node 0, the parent node number information is given as blank because Node 0 is the root node of decision tree. Variable name has been mentioned as next column, which provides the name of the independent variable predicted at every node. Node 0 of the decision tree model depicts the target variable or dependent variable named as victim_minor with two categories Y and N. That is why the variable name column of Table 3 is blank for Node 0. Next column has been provided Chi-Square test value. Chi-Square value has been calculated for every node at the time of the classification. Node 0 depicts the target variable so the Chi-Square value has not shown for this particular node. This value has been measured for further nodes to

check the significance level of each node. For Nodes 1 and 2, Chi-Square value=172.187, Chi-Square value found 23.156 for Nodes 3, 4, 5 and for Nodes 6 and 7, the measurement of test is 22.324. The last column of the table indicates significant value, which is less than .05 for every node. Total depth of the decision tree has been found as 2. So, Node 1 has been found the parent of Nodes 3, 4, 5 and Node 2 has been found as parent of Nodes 6 and 7. Very less number of cases of victim age 'Major' classified at Node 3 with the rate of 9.2%. At Node 7, highest number of cases found of victim age 'Major' with percentage of 91.2%. In case of victim age 'Minor' category, Node 3 classified the highest number of cases with percentage value 90.8% and the lowest percentage occurred at Node 7 for the classification of victim minor. Nodes 3 and 7 has been found as terminal nodes of decision tree model. Predicted category column of Table 3 indicated category wise prediction at every node. For example: Y has been predicted at Nodes 0, 1, 3, 4, 5 and N has been predicted at Nodes 2, 6 and 7. As Table 3, Table 4 has also been sown the information about each node of decision model which has been formed by 30% test data of the entire data set shown in Figure 2. All depicted nodes are same with same types of independent variables but the values are different. Here, minimum values are found at Node 3 with 11.5% for major category and at Node 7 for minor category with 15.2%. Maximum values of victim age 'Major' category has been found at Node 7 with 84.8% and victim age 'Minor' found at Node 4 with 88.5% with highest rate.

Table 3 Decision tree detailed table from training data using CHAID

Node No.	*Cases in major*	*Percentage*	*Cases in minor*	*Percentage*	*Predicted category*	*Parent node*	*Variable name*	*Chi-Square value*	*Significance*
0	208	40.9%	301	59.1%	Y				.000
1	51	17.0%	249	83.0%	Y	0	RACE	172.187	.000
2	157	75.1%	52	24.9%	N	0	RACE	172.187	.000
3	9	9.2%	89	90.8%	Y	1	AGE	23.156	.000
4	23	37.1%	39	62.9%	Y	1	AGE	23.156	.000
5	19	13.6%	121	86.4%	Y	1	AGE	23.156	.000
6	74	62.7%	44	37.3%	N	2	WEIGHT	22.324	.000
7	83	91.2%	8	8.8%	N	2	WEIGHT	22.324	.000

Notes: When Race= White Hispanic, white Asian /pacific island, black Hispanic and Age<=39 then minor victim=90.8% and major victim=9.2%.

When Race=White Hispanic, white Asian /pacific island, black Hispanic and 39<Age<=48 then minor victim=62.9% and major victim=37.1%.

When Race=White Hispanic, white Asian /pacific island, black Hispanic and Age>48, then minor victim=86.4% and major victim=13.6%.

When Race=Black and Weight <=190, then minor victim=37.3% and major victim=62.7%.

When Race= Black and Weight > 190, then minor victim=8.8% and major victim=91.2%.

Table 4 Decision tree detailed table from test data using CHAID

Node No.	*Cases in major*	*Percentage*	*Cases in minor*	*Percentage*	*Predicted category*	*Parent node*	*Variable name*	*Chi-Square value*	*Significance*
0	96	42.7%	129	57.3%	Y				.000
1	24	18.5%	106	81.5%	Y	0	RACE	172.187	.000
2	72	75.8%	23	24.2%	N	0	RACE	172.187	.000
3	12	29.3%	29	70.7%	Y	1	AGE	23.156	.000
4	3	11.5%	23	88.5%	Y	1	AGE	23.156	.000
5	9	14.3%	54	85.7%	Y	1	AGE	23.156	.000
6	33	67.3%	16	32.7%	N	2	WEIGHT	22.324	.000
7	39	84.8%	7	15.2%	N	2	WEIGHT	22.324	.000

On the same data set Exhaustive CHAID has been applied to compare accuracy and prediction rates of trained and tested models. Here, also in Exhaustive CHAID the independent variable Race has been found as the best predictor. The decision tree diagram has been formed in Figure 3. Age has been found as second good predictor, which has been categorised into two values i.e. Age<=45 and Age>45 when Race=Black. The following rules have been generated from the Figure 3 and Table 5.

Table 5 Decision tree detailed table from training data using exhaustive CHAID

Node No.	*Cases in major*	*Percentage*	*Cases in minor*	*Percentage*	*Predicted category*	*Parent node*	*Variable name*	*Chi-Square value*	*Significance*
0	217	41.6%	305	58.4%	Y				
1	52	17.1%	252	82.9%	Y	0	RACE	179.383	.000
2	165	75.7%	53	24.3%	N	0	RACE	179.383	.000
3	48	60.8%	31	39.2%	N	2	AGE	15.006	.005
4	117	84.2%	22	15.8%	N	2	AGE	15.006	.005

Notes: When Race= Black and Age<=45 the predict minor victim=39.2% and major victim=60.8%

When Race= Black and Age>45 the predict minor victim=15.8% and major victim=84.2%

The classification and prediction of the sexual offenders have been modelled in Figure 3 through the training data using Exhaustive CHAID. The detailed information of the model has been depicted as in Tables 5 and 6, separately for training and test data. The Chi-Square value of every node has been given. The entire tree diagram has been formed through 5 nodes.

Table 5 depicts the detailed summary of every node of decision tree model which has been deployed with 70% of trained data using Exhaustive CHAID algorithm as explained by Figure 3. The detailed information of 5 nodes have been provided in this table because only two independent variables are included here to form decision model as predictors

i.e. Age and Race. All nodes have been found with their significant values which are less than 0.05. Node 0 indicates the root node which has been also found as the parent node for Nodes 1 and 2 and Node 2 depict the parent node of Nodes 3 and 4. Here in Table 5, lowest rate of 'Major' category occurred at Node 1 with 17.1% and lowest rate of 'Minor' category occurred at Node 4 with 15.8%. The highest percentage achieved at Node 4 for major category with 84.2% and at Node 1 for minor category with 82.9%. Table 6 also explained the detailed information of the decision tree model created through the Exhaustive CHAID method with rest of 30% data set of the entire data set. The chi-square value has been found as same here as in Table 5 for all the terminal nodes.

Table 6 Decision tree detailed table from test data using exhaustive CHAID

Node No.	*Cases in major*	*Percentage*	*Cases in minor*	*Percentage*	*Predicted category*	*Parent node*	*Variable name*	*Chi-Square value*	*Significance*
0	87	41.0%	125	59.0%	Y				
1	23	18.3%	103	81.7%	Y	0	RACE	179.383	.000
2	64	74.4%	22	25.6%	N	0	RACE	179.383	.000
3	22	68.8%	10	31.3%	N	2	AGE	15.006	.005
4	42	77.8%	12	22.2%	N	2	AGE	15.006	.005

4 Performance measure

The performance of the decision tree model has been measured by the Gain and Index percentages and their charts separately, both for CHAID and Exhaustive CHAID models. As shown in Table 7, all 5 terminal nodes have been depicted with total number of cases occurred in both the categories i.e. minor and major, as per the trained and tested models. Total percentage column indicates the value (in percentage) of total number of cases occurred in both the categories of target variable. Percentage column indicates the percentage of number of cases predicted by the particular node, divided by total number of cases. For Example: for terminal Node 7, 91 cases have been predicted through this node and total number of cases are 509, therefore calculated percentage is17.9%.

Figure 5 shows the gain chart for both the trained and tested model using CHAID method. Gain value provides the entire information for the major category because major category has been set as target category for the proposed model. During the development of the decision tree model the target category of dependent variable has been chosen as 'N' i.e. major category. So, Gain Node column has shown the total number of cases occurred in the target category. As per the data set 83 cases can be correctly predicted by the decision model using CHAID for major category (when race=black and weight >190.0) because the target category is defined as major category.

Table 7 Performance measures calculation of CHAID Model in terms of gain, index and response for training and test data set

Terminal node	*Total cases*	*Total percentage*	*Gain node (N)*	*Gain per cent*	*Response per cent*	*Index per cent*
For training sample						
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%
For test sample						
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%

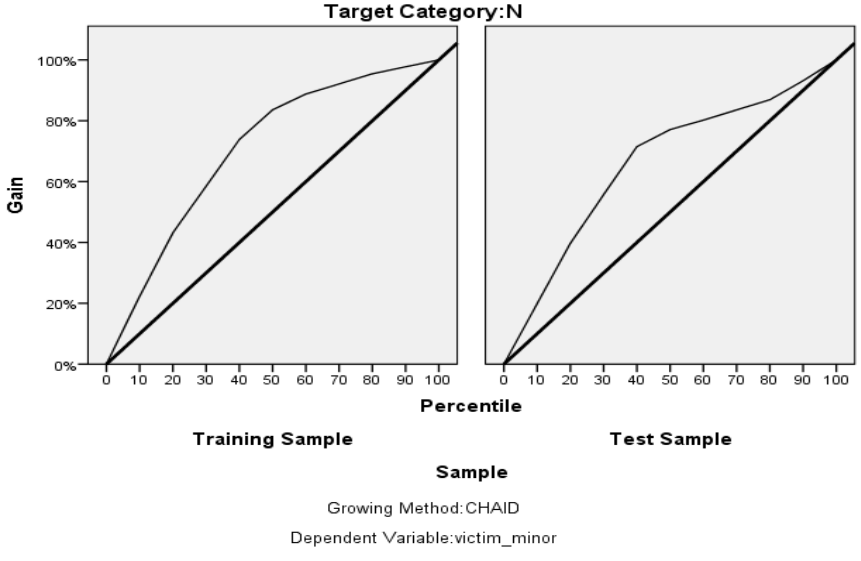
Notes: Gain has been defined by the total number of cases in the target category for each terminal node computed by the formula as:

(node: number of target cases (*n*) / total number of target cases)*100

Gain chart is simple line graph computed through Cumulative Percentile (CP) as:

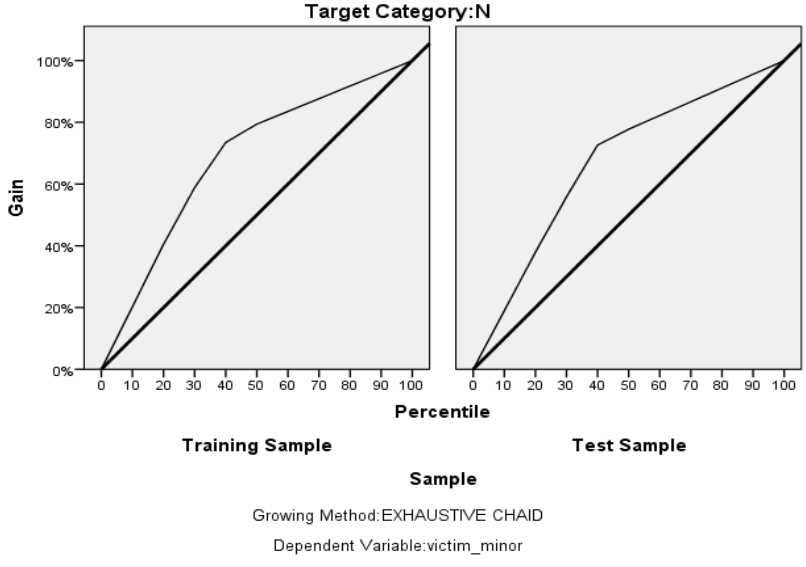
(CP of number of target cases/ total number of target *n*)*100.

Figure 5 Gain chart for training and test data (CHAID)



As depicted in Table 7, Response value is defined by the total percentage of the target category i.e. major victim, defined in the trained and tested model using CHAID. As shown in the Table 7, 91.2% responses have been depicted for terminal 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. For terminal Node 6, 62.7% sexual offenders can be classified on the basis of target category i.e. major category when race of the sexual offender = black and weight <= 190. 37.1% response percentage has been calculated at Node 4 of decision model, when race of sexual offenders is found from white Hispanic; white; Asian/Pacific; black Hispanic and Age is between 39 and 48. At next terminal Node 5 the response rate occurred as 13.6%, which means that only 19 records of sexual offenders are classified at this particular node under its target category and when race has been considered as white Hispanic; white; Asian/Pacific; black Hispanic and Age is > 48. Response percentage at last terminal Node 3 has been found as 9.2% which means that only 9 records of sexual offenders are found under major category. Performance measures of test Sample are provided almost similar results as performance measures of training sample as shown in Table 7. Trained model has been validated on the basis of tested model. As per the performance measurement through gain percentage, it has been observed that a very small change has been found in the training and test data except two nodes. Node 4 has been found with the huge difference between the performance measurements of training and test data. For training sample it depicts 11.1% and for test it has been reduced to 3.1%. Similar difference between values have been found in case of Node 3, for training data the gain percentage shows 4.3% and for test it has increased till 12.5%.

Figure 6 Gain chart for training and test data (exhaustive CHAID)



Index has been shown as last column of Table 7, 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 same. Index

chart provides the information about the goodness of model. For good models index chart must start above 100%. As depicted in Figure 9, the index chart has been started above 220% in both the trained and tested model, using CHAID growing method. Figure 10 provides the Index chart using Exhaustive CHAID algorithm, which has been started few points above 200%. Both the charts have been depicted the goodness of models.

Figure 7 Response charts for training and test data (CHAID)

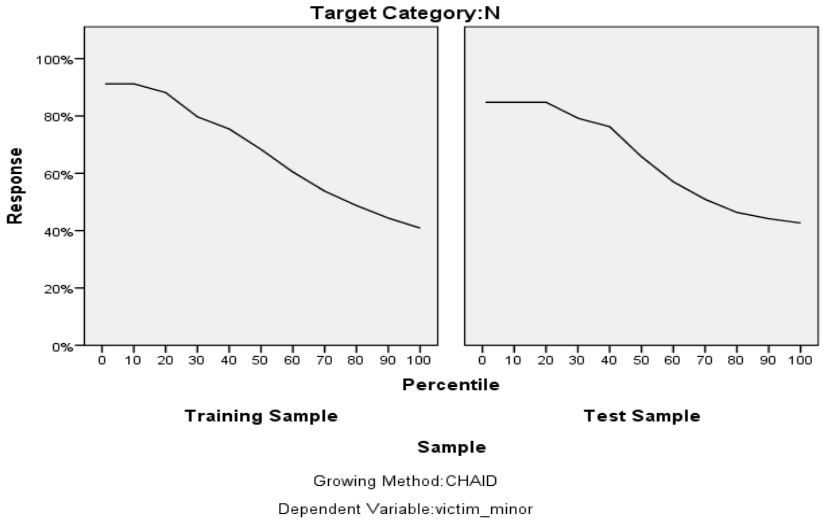


Figure 8 Response charts for training and test data (exhaustive CHAID)

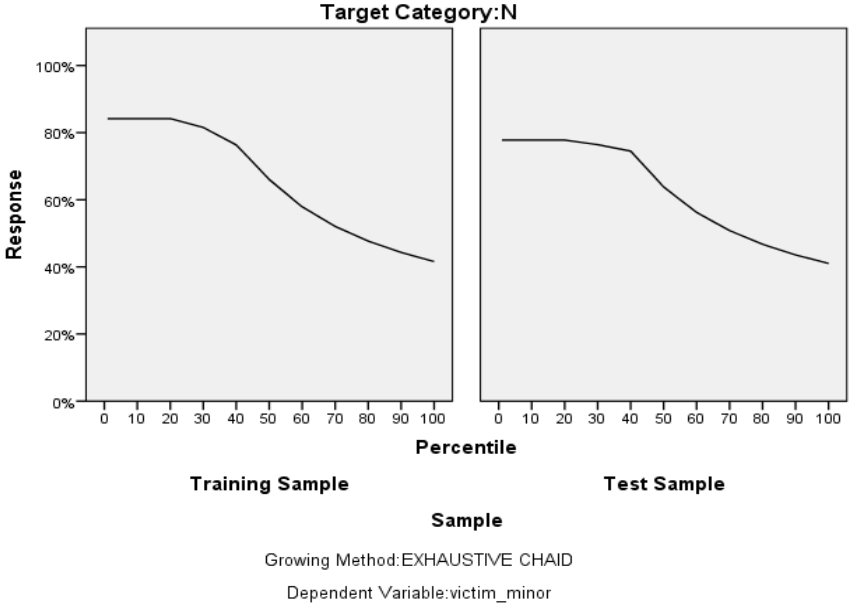


Figure 9 Index charts for training and test data (CHAID)

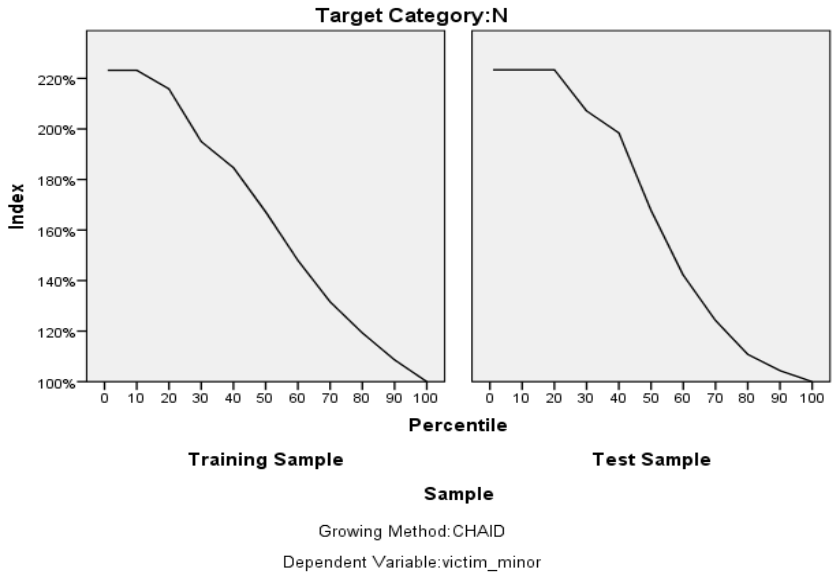


Figure 10 Index charts for training and test data (exhaustive CHAID)

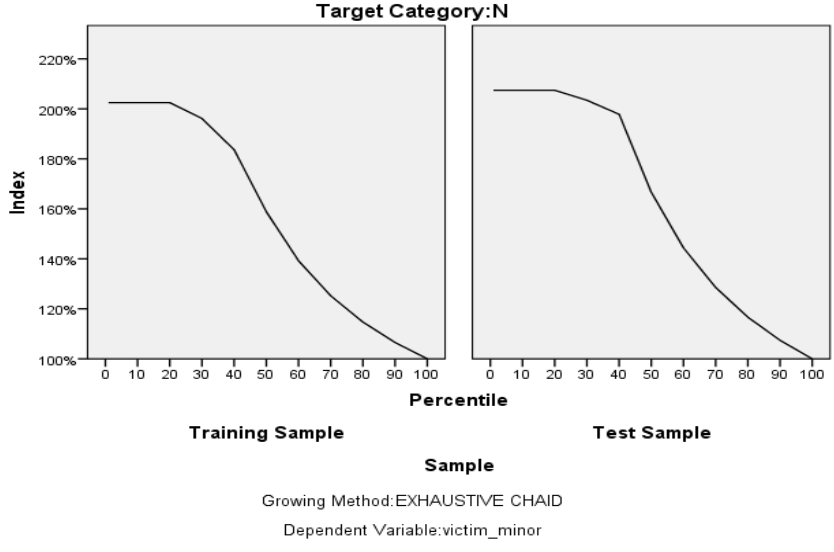


Table 8 depicts the performance measures of the decision model developed by Exhaustive CHAID algorithm. Even the performance measures of CHAID have been shown by Table 7. By exhaustive CHAID model only 3 terminal nodes have been formed by the trained and tested decision tree as shown in the Figures 3 and 4. So, various information of these 3 nodes (4, 3, 1) has been given in terms of gain, response and index

percentages. Index and gain charts have also been developed to depict the performance of the exhaustive CHAID model as shown by Figures 6 and 10. The percentile values have created the line graph in gain and index charts. At Node 2, gain percentage is very low as compared to the other nodes because only 48 cases have been occurred with major category. 24% gain has been counted at Node 1 where 52 cases under major category are found. The highest gain percentage is 53.9% which has been measured at Node 4 with the total number of 117 cases of major category. As resulted by the test sample 48.3% gain measured at Node 4, 26.4% and 25.3% at Nodes 1 and 3. So the lowest gain has been found at Node 3 in both test and training sample data. The response percentages are also higher at Node 4 for training and test data as gain percentage i.e. 84.2% and 77.8%. Response percentages of Node 1 are low i.e. 17.1% and 18.3%. Their respective charts are also shown in Figure 8 for training and test sample using exhaustive method of growth. The index percentage is 202.5% under Node 4 through training data and 189.5% through test data.

Table 8 Performance measures calculation of exhaustive CHAID model in terms of gain, index and response for training and test data set

Terminal node	*Total cases*	*Total percentage*	*Gain node (N)*	*Gain per cent*	*Response per cent*	*Index per cent*
For training sample						
4	139	26.6%	117	53.9%	84.2%	202.5%
3	79	15.1%	48	22.1%	60.8%	146.2%
1	304	58.2%	52	24.0%	17.1%	41.1%
For test sample						
4	54	25.5%	42	48.3%	77.8%	189.5%
3	32	15.1%	22	25.3%	68.8%	167.5%
1	126	59.4%	23	26.4%	18.3%	44.5%

Table 9 has shown the risk estimation for both the developed models (CHAID and Exhaustive CHAID). It has been found that 20.2% risk is there for using CHAID decision model to predict the sexual offenders and 20.1% risk found by Exhaustive CHAID decision model. Risk has been validated through the test models where 20.9% and 21.2% have been found by 30% of test data through both the techniques CHAID and Exhaustive CHAID. Few fraction of error can be occurred in both the models i.e. .018 and .019.

Table 9 Risk estimation through CHAID and exhaustive CHAID

Sample	*Training (CHAID)*	*Training (Exhaustive CHAID)*	*Test (CHAID)*	*Test (exhaustive CHAID)*
Risk estimation	.202	.201	.209	.212
Error	.018	.019	.027	.028

Developed models also provide the predicted accuracy results as shown in Table 10. So, there could be approximate 20% chances of wrong prediction but accuracy rate is approximately 80% as measured through both the methods CHAID and Exhaustive CHAID. Table has shown the predicted accuracy values in overall four cases. (i) Accuracy through CHAID trained model (ii) Accuracy through Exhaustive CHAID

trained model (iii) Accuracy trough CHAID test model (iv) Accuracy through Exhaustive CHAID test model. First of all in case (i), predictive accuracy results of CHAID training model has been shown in terms of N and Y. Here, N means the predicted accuracy rate of major age group and Y means the predicted rate of accuracy for minor age group. Through this model the overall prediction rate of target category or major category has been found as 41.1% and for minor category it has measured 58.9%. Total percentage of correctness through model found as 79.8% which is approximately equal to 80%. As resulted, this is the very good percentage to predict the victim's age. As per the individual categories prediction, the overall percentage of correctness for major victim is 75.5% and 82.7% is for minor. Second case has been provided the predictive accuracy rates for tested model. CHAID model has been validated through test model because the overall rate of correctness is almost equal with the rate of correctness of trained model. 79.1% of accuracy rate has been found through validation. Third case has shown the results of correctness to predict the major and minor age groups through Exhaustive CHAID model. The results of accuracy prediction are almost same as through CHAID and Exhaustive CHAID. Exhaustive CHAID model can predict major victim age with 76% rate of accurateness and can predict minor victim age with 82.6% of accuracy. The overall rate of accuracy or correct prediction is 79.9% through Exhaustive model. Exhaustive model has also been validated and results have been generated for rate of correctness. So, exhaustive decision model is working correctly because the prediction rate of test model is almost same as the prediction rate of trained model. The results of both the CHAID and Exhaustive CHAID results have been validated with very good accuracy rate.

Table 10 Predicted accuracy results through models for training and validation

Sample/Technique	*Observed*	*Predicted-major(N)*	*Predicted-minor(Y)*	*Per cent correct*
Sample-training (CHAID)	N	157	51	75.5%
	Y	52	249	82.7%
	Overall percentage	41.1%	58.9%	79.8%
Sample-test (CHAID)	N	72	24	75.0%
	Y	23	106	82.2%
	Overall percentage	41.1%	58.9%	79.1%
Sample-training (exhaustive CHAID)	N	165	52	76.0%
	Y	53	252	82.6%
	Overall percentage	41.8%	58.2%	79.9%
Sample-test (exhaustive CHAID)	N	64	23	73.6%
	Y	22	103	82.4%
	Overall percentage	40.6%	59.4%	78.8%

5 Conclusions

This research paper has proposed decision models to predict the sexual offenders on the basis of minor and major victim age groups by using CHAID and Exhaustive CHAID growing methods. First CHAID has been used on dataset then Exhaustive CHAID has

also being applied to check the accuracy of results. Good accuracy rates have been derived by both the decision models to predict the sexual offenders of victims under different age groups (major and minor). The performance of decision models has been measured through gain percentages and gain charts, response percentages and response charts further index percentages and index charts for both trained and tested models. It has been found that both decision tree models have provided very good results for the classification and prediction of sexual offenders for different victim age groups. 70% of the entire dataset has been used to develop the model and 30% data set has been used to validate the decision models. As found from the results through CHAID and Exhaustive CHAID models, the overall rates of accuracy are 79.8% and 79.1%to predict the cases correctly. It has been found that approximately 80% results will be correct while using these models and approximately 20% of risk can be evaluated. The goodness of the model has also been validated through the performances charts.

It has been clear that both the models provided almost same value of accuracy. Further, proposed models have also been validated through remaining 30% of the data set. Model validated successfully because through the test data overall 79.1% rate of accuracy has been delivered using CHAID, which has been found very close to the accuracy rate provided by trained model of CHAID method. To predict the cases 78.8% rate of accuracy has been found through the validation of Exhaustive CHAID trained model. The trained model measured 79.9% accurate rate of prediction. So, it has been interpreted that both the models are validated with correct results. For the development of the decision tree models using CHAID and Exhaustive CHAID, total 4 variables i.e. age, weight, height, race were involved initially but CHAID considers only 3 variables as good predictors (race, age, weight) for the development of decision tree model. Whereas Exhaustive CHAID considers only two independent variables as predictors for the model development i.e. race and age. These proposed models can be used by the police departments and various law organisations to classify and predict the sexual offenders for major and minor age groups so that immediate actions can be taken accordingly.

References

Ahmed, M.D., van Niekerk, R.L., Ho, W.K.Y., Morris, T., Baker, T., Ali Khan, B. and 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*, Vol. 42, No. 1, pp.33–53.

Barth, J., Bermetz, L., Heim, E., Trelle, S. and Tonia, T. (2013) 'The current prevalence of child sexual abuse worldwide: a systematic review and meta-analysis', *International Journal of Public Health*, Vol. 58, No. 3, pp.469–483.

Biggs, D., De Ville, B. and Suen, E. (1991) 'A method of choosing multiway partitions for classification and decision trees', *Journal of Applied Statistics*, Vol. 18, No. 1, pp.49–62.

Browne, K.D., Hines, M. and Tully, R.J. (2018) 'The differences between sex offenders who victimise older women and sex offenders who offend against children', *Aging and Mental Health*, Vol. 22, No. 1, pp.11–18.

Goodman, L.A. (1979) 'Simple models for the analysis of association in cross classifications having ordered categories', *Journal of the American Statistical Association*, Vol. 74, pp.537–552.

Kaur, B., Ahuja, L. and Kumar, V. (2018) 'Factors affecting crime against women using regression and k-means clustering techniques', *Industry Interactive Innovations in Science, Engineering and Technology*, Springer, Singapore, pp.149–162.

Kaur, B., Ahuja, L. and Kumar, V. (2020) 'Modeling the factors affecting crime against women: using ISM technique', *Advances in Computing and Intelligent Systems*, Springer, Singapore, pp.469–478.

Kass, G.V. (1980) 'An exploratory technique for investigating large quantities of categorical data', *Applied Statistics*, Vol. 20, No. 2, pp.119–127.

McDonald, P. (2012) 'Workplace sexual harassment 30 years on: a review of the literature', *International Journal of Management Reviews*, Vol. 14, No. 1, pp.1–17.

Marshall, W.L., Barbaree, H.E. and 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*, Vol. 18, No. 4, pp.11939–11944.

Oh, E.S., Choi, J.H., Lee, J.W. and Park, S.Y. (2018) 'Predictors of intentional intoxication using decision tree modeling analysis: a retrospective study', *Clinical and Experimental Emergency Medicine*, Vol. 5, No. 4, pp.230–239.

Quist, J., Mirza, H., Cheang, M.C.U., Telli, M.L., Lord, C.J., Tutt, A.N.J. and Grigoriadis, A. (2017) 'Association of a four-gene decision tree signature with response to platinum-based chemotherapy in patients with triple negative breast cancer', *Molecular Cancer Therapeutics*, Vol. 18, No. 1, pp.204–212.

Upadhyaya, A., Shokeen, V. and Srivastava, G. (2018) 'Decision tree model for classification of fake and genuine banknotes using SPSS', *World Review of Entrepreneurship, Management and Sustainable Development*, Vol. 14, No. 6, pp.683–693.

WHO (2014) *Global and regional estimates of violence against women*, World Health Organization, 28 November 2014. Available online at: https://www.who.int/reproductivehealth/publications violence/9789241564625/en/