An Approach Towards Development of a Predictive Model for Female Kidnapping in India Using R Programming

Sumit Chatterjee, Surojit Das, Sourav Banerjee and Utpal Biswas

Abstract The major concern of the present world is the increase in criminal activities that are taking place throughout the world. The criminal activities in today's world include murder, theft, rape, women exploitation, human trafficking, possession of illegal properties, kidnapping. This paper summarizes the criminal activities related to female kidnapping in India. This paper highlights the statistical analysis of female kidnapping in India and thereby develops a predictive model to envisage the purpose of kidnapping of an individual female based on certain parameters. The model is developed using the decision tree technique by applying Iterative Dichotomizer (ID3) algorithm. The ID3 algorithm uses the entropy measure as a criterion for selecting classifiers for branching. The raw data set is converted to an appropriate one by converting the categorical values to numerical values using label and one hot encoding. This model is then trained with the appropriate training data set, and then, its performance is evaluated with a testing data set. The efficiency of the model is detected with the measures of accuracy, precision, recall, F1, AUC scores.

Keywords Human trafficking · Kidnapping · Statistical analysis · Decision tree ID3 algorithm · Entropy · Classifier · Label and one hot encoding · Accuracy Precision · Recall · F1 · AUC

S. Chatterjee (✉) · S. Banerjee
Kalyani Government Engineering College, Kalyani, Nadia 741235, India
e-mail: sumitc219@rediffmail.com

S. Banerjee
e-mail: mr.sourav.banerjee@ieee.org

S. Das
Brainware University, Kolkata 700124, West Bengal, India
e-mail: surojitdas90@gmail.com

U. Biswas
University of Kalyani, Kalyani, Nadia 741235, India
e-mail: utpal01in@yahoo.com

M. Chakraborty et al. (eds.), *Proceedings of International Ethical Hacking Conference 2018*, Advances in Intelligent Systems and Computing 811,
https://doi.org/10.1007/978-981-13-1544-2_40

1 Introduction

The crime is not a problem of an individual, but it is a problem of the entire society [1]. It influences the socio, political and economic aspects of a country. The various forms of crime in India are murder, theft, women exploitation, human trafficking, prostitution, etc., [1]. The social problems can occur due to unpredictable social changes and unwillingness to adapt in new social environment [1]. The rapid changes in social environment often trigger crime. Though different societies have different changes, they may stimulate the same kind of crime rates [1]. The women in India are victims of crime, exploitation and violence. According to the statistical report of National Crime Record Bureau in 2015, in every eighth minute a woman is being kidnapped in India [2]. The criminal activities are also spread using illicit Web domain (dark nets) and social networks [3]. The websites of dark nets provide advertisements which are actually traps for women trafficking [3]. The women are kidnapped for various purposes like adoption, begging, marriage, illicit intercourse, camel racing, prostitution, ransom. This paper has focussed on the evaluation of the statistics relating to the mentioned purposes of kidnapping. The women trafficking are ranked third in organized crime list. According to the Human Rights Commission of India, every year 40,000 children are kidnapped and 11,000 remain untraced [4, 5].

2 Related Work

The crime analysis using statistical tools and machine learning algorithms is an interesting area of research nowadays [6]. But not too many research works have been done in this field. There have been works done to identify behavioural patterns related to human trafficking by extracting information from newspaper, social networks and illicit Web domains and generating corpus vector containing details about human trafficking [7, 8]. The techniques which were used were data mining for finding the hidden information, data pre-processing to gather all the information, parsing for obtaining a semantic meaning for the collected information, using techniques like deep learning [9] and NLP [9] to classify information and find the words related to trafficking. In another project [10], a website called ‘Backpage’ which is used for classified advertisement has been used as a source of information for many of the research works [11, 12]. The machine learning algorithms (semi-supervised) [10] are used to generate predictive model which is trained with labelled and unlabelled data and then tested on unseen data to evaluate their efficiency. ‘Backpage’ has also been used by a research team of undergraduate students for examining sex traffic in Pennsylvania [11–13]. The prostitution advertisements of different countries were analysed, and the major countries involved in trafficking were identified. Females involved were identified and formed three subnetworks, namely disconnected subnetworks, high-density interconnected subnetworks and sparsely interconnected subnetworks [13].

Open-source data available [14] from the Internet have been also used to design a framework for indicating human trafficking. A three-level model has been proposed by Glynn Rankin [14] to understand and report all the illegal activities related to trafficking across the network. The level 1 deals with the credible and accepted indicators, the level 2 describes the materials produced as a result of anti-trafficking activities, and the level 3 deals with social media and citizen-created content. This model was developed in challenging Boisot's knowledge management model [15] or iSpace [15] which is a three-dimensional cube-like structure.

In a project of domestic human trafficking, movements along the circuits (systematic movement of provider to various cities for promoting sex activity) have been given stress [16, 17]. The network analysis methods [16] were used to find the circuits from GIS data. Various escort advertisements were collected with a count of about 90 advertisements per day, duplicate advertisements were eliminated, and then, the location of each advertisement was analysed to identify intra-state circuit. In another paper [18] related to crime analysis in Haryana, a spatio-temporal analysis of abduction of women in the city of Chandigarh has been carried out to find out the areal variation and distribution. The factors like socio-culture, economic condition and population explosion have been identified as one of the major causes for such crimes. This project also proposes [18] a formula:

$$Crime\ Rate = Cri = (CXi/TFPi) * 100{,}000 \tag{1}$$

where

CRi rate of crime 'X' in wards I,
CXi crime 'X' in wards I,
TFPi total female population in wards I

Another research paper [19] says that the sex traffickers target children more than the adults. Thus, an initiative can be taken to prevent child trafficking by employing a child safety system with the use of smart identity card. The smart identity card is enabled with radio frequency identification device (RFID) which can help in tagging the school children. The objective of this paper was to ensure security by reducing the gap between parents, children and teachers. A mobile safety monitoring system [20] was proposed in this paper that informs the guardians about the information relating to the safety of their children. A school security system (SSS) [21] prototype model was also proposed based on Web-based development using PHP, VB.net, Apache Web server and MySQL. The out time and in time of each student will be recorded successfully in the system, and the GPS technology will enable the parents to receive continuous information about their children.

Sometimes, analysing the advertisements of the sex workers can too work. The rigorous content analysis of online advertisements of sex workers can provide certain behavioural patterns that can identify the victims of human trafficking [22]. Statistics suggest that about 75% of the advertisements possess more than one primary indicator. The virtual indicators that are discovered are: movement, shared management, controlled/restricted movement, advertisements posted by third party, advertisement ethnicity/nationality. To provide answers to entry-centric questions on human traffic data to help the investigators is a matter of challenge. An entity-centric knowledge graph can be developed to build a semantic search engine to help the investigators [23]. The idea is to take a large number of advertisements from the Web and create a knowledge graph. This allows the investigators to search their queries with a well-built semantic search engine.

3 Proposed Work

3.1 Statistical Analysis of the Female Kidnapping Records in India

The paper uses a data set which contains the records of kidnapping of women of different ages for various purposes. Based on these records, a statistical analysis has been performed to find out the number of females kidnapped of different age categories and for different purposes like begging, prostitution.

3.2 Design of a Predictive Model for Female Kidnapping

Certain parameters have been identified which can act as classifiers for designing a predictive model for female kidnapping. These parameters are age, financial status, glamour and marital status. A predictive model has been designed using these parameters as classifiers by constructing a decision tree [24]. A decision tree is a supervised machine learning technique used for solving classification problems. It can work with both categorical and continuous data and is widely used for multilevel classification problems. In a decision tree, the population is split into two subpopulations based on the significant classifier (parameter). The algorithm which is used to construct the decision tree is Iterative Dichotomizer (ID3) [25] algorithm. It uses a top-down approach to construct the tree. It is based on maximizing the information gain and minimizing the entropy [26]. At each node, every possible property is tried and the property which maximizes the information gain and minimizes the entropy is selected to split the node. This process is recursively repeated until all the leaf nodes are homogeneous (contains objects of same classes). It is a greedy algorithm, always uses entropy as a criterion, and never looks for alternative choices like Gini index, classification error. The predictive model can predict for what purpose (begging, prostitution, etc.) an individual has been kidnapped based on these parameters.

Figure 1 shows the decision tree of classification of iris data set. Iris data set contains the features of three classes (species) of flowers, namely setosa, versicolor and virginica. Four features have been specified in the data set, namely sepal length, sepal width, petal length and petal width. The data set has 150 samples with 50

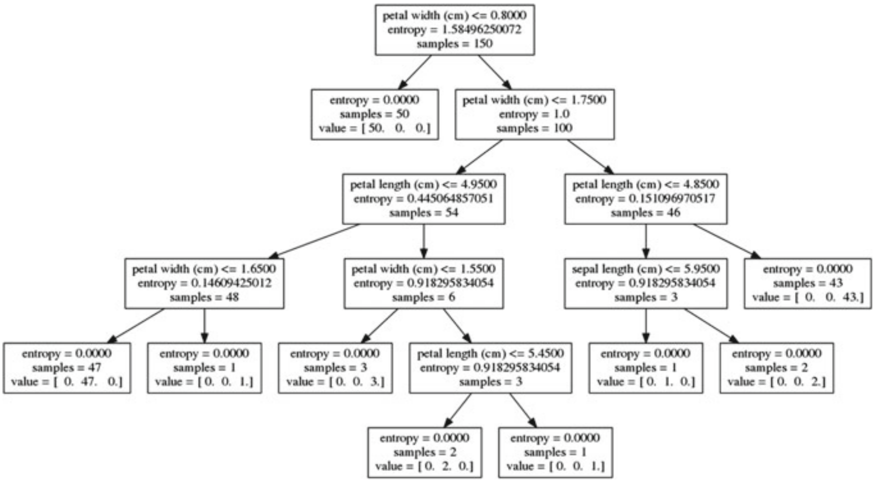


Fig. 1 An Illustration of decision tree for classification of iris data set [27]

samples belonging to each of the classes. The root node selects petal width as one of the classifiers based on the value of entropy. At each step, the decision tree classifies the samples into different classes by selecting different classifiers with a motive to minimize the entropy. The leaf nodes denote the final decision regarding the classification. This kind of approach has also been used in the decision of the proposed model discussed later in the Sect. 4.3.

All the related works which have been mentioned in Sect. 2 have been done on the data available from the Internet. This paper uses the manual data obtained from the records of the Government of India [28] and hence covers the cases which include all kinds of kidnapping and trafficking of women including those via the Internet. Thus, the scope of applicability of this paper is comparatively more because it deals with larger data set. Moreover, the proposed model also predicts the purposes of kidnapping which can help the police to investigate and trace the criminals who deal with that particular purpose. Thus, this paper has a unique importance in comparison with the papers mentioned in the related works. The predictive model that has been designed is also beneficial to the society as police can easily search among the criminals involved in a specific purpose for kidnapping thereby enhancing the investigation process and reducing the time.

4 Experimental Results

4.1 Discussion on Environment Setup

The coding for the proposed work has been implemented using R programming. R is a programming language that is widely used by mathematicians and scientists for statistical computing. It is very popular among statisticians and data miners and also supports graphical works. The IDE which has been used for the implementation of the proposed work is R Studio 3.4.1. It is an open-source software developed by the Comprehensive R Archive Network (CRAN) and available for platforms like Windows, Linux, MAC [29]. The decision tree which has been used to create the predictive model can be viewed in R Studio 3.4.1, but for better visualization another software Graphviz 2.83 is used. It is a third-party software and is used widely for the visualization of complicated and large-sized images.

4.2 Data Set Used

The data set provides the information that among the major purposes behind female kidnapping, the most predominant purposes are: begging, marriage, illicit intercourse, prostitution, ransom.

Volume of Data Set. Number of samples in training data = 150, number of samples in testing data = 50.

Clustering. Clustering is done based on the purposes of kidnapping. The clusters and their size are specified in Table 1.

The parameters and their ranges which have been selected to predict the purpose behind kidnapping are mentioned in Table 2.

The categorical data like financial status, glamour, marital status are converted to numerical values using label encoding [30]. The label encoding is used to transform non-numeric values into numeric values. The numeric values after encoding range from 0 to number of classes—1. The conventions used are mentioned in Table 3.

Table 1 Clustering of female kidnapping data

Cluster (class)	Number of samples in each cluster (class)
Begging	30
Marriage	30
Illicit intercourse	30
Prostitution	30
Ransom	30

Next one hot encoding or dummy encoding is used to generate new columns [30]. The name ‘dummy’ suggests the use of a duplicate variable to represent one level of categorical variable. If a level is present, then it is indicated by ‘1’, and if a level is absent, then it is indicated by ‘0’. One dummy variable is created for each and every level. The application of one hot encoding on the data generates columns having binary values 0 and 1 for each of the kidnapping types/causes. Suppose begging is a cause for kidnapping. So a new column called begging is generated which has values of 1 for all the records (rows) which correspond to begging and 0 for the others. The similar work is done for other causes like marriage, illicit intercourse, prostitution, ransom.

4.3 Construction and Analysis of Decision Tree

A predictive model is created by analysing the pre-processed data set using a decision tree.

The predictive model designed using decision tree shown in Fig. 2 is trained using 150 sample records. The decision tree starts initially with 150 samples. Then, it has to choose a parameter as classifier for classifying the samples into specific categories. The parameter is chosen on the basis of the entropy value [31]. Entropy is the measure of impurity or uncertainty. So higher values of entropy mean more uncertainty, while lower entropy values mean less uncertainty. Thus, the target of the decision tree is to

Table 2 Data set of female kidnapping

Age	Below 10	10–15	15–18	18–30	30–50	Above 50
Financial status	Below 1 lakh		1–5 lakh		Above 5 lakh	
Glamour	Low		Medium		High	
Marital status	Unmarried		Married			

Table 3 Label encoding of categorical data

Parameters	Categorical values	Numerical values
Financial Status	Below 1 lakh	1
	1–5 lakh	2
	Above 5 lakh	3
Glamour	Low	1
	Medium	2
	High	3
Marital status	Unmarried	0
	Married	1

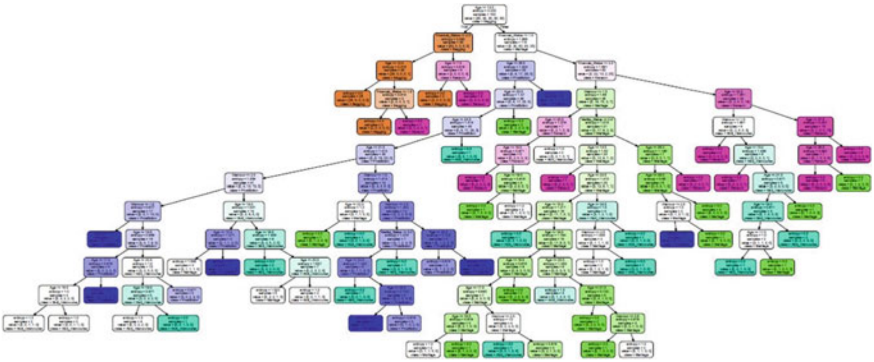


Fig. 2 Decision tree classification model of female kidnapping

minimize the value of entropy as far as possible. The minimum value of entropy is 0 which means that this condition has no uncertainty; i.e., it is certain.

The root node contains 150 sample records with 30 records for each of the kidnapping purposes (begging, marriage, illicit intercourse, prostitution, ransom). The class begging is chosen as label of the node arbitrarily as all of the classes (kidnapping purposes) has equal number of sample records; i.e., each has 30 sample records. The parameter age is chosen as the classifier with an entropy value of 2.322. The threshold value of age is selected as 13.5. At the next level of the tree, two child nodes are formed from the root node. One of the child nodes contains sample records with age values less than or equal to 13.5, and the other child node contains sample records with age values greater than 13.5. The class begging is chosen as label of the first child node as the majority number of records correspond to begging. The first child node has an entropy value of 0.592, and financial status is chosen as the classifier for the next level with a threshold value of 2.5. The class marriage is chosen as label of the second child node as the majority number of records correspond to marriage. The second child node has an entropy value of 1.996, and financial status is chosen as the classifier for the next level with a threshold value of 1.5. This process of generation of child nodes continues until the leaf nodes containing sample records belonging to a particular category (kidnapping purpose) are generated. The model is subjected to a testing set of data with 50 samples, and its efficiency of classification is measured based on the following scores:

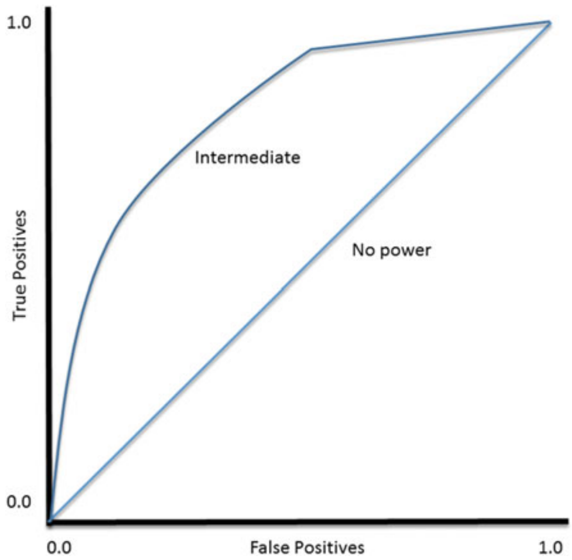
Accuracy. It is defined as the number of correct predictions out of the total number of predictions [32].

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{2}$$

Precision. It is defined as the number of actual true cases that has been detected out of the total number of predicted true cases [32].

$$Precision = TP/(TP + FP) \tag{3}$$

Fig. 3 ROC curve [33]



Recall. It is defined as the number of true cases which has been predicted correctly out of the total number of true cases [32]

$$Recall = TP/(TP + FN) \quad (4)$$

F1 Score. It is the harmonic average for precision and recall. It has a value of 1 for ideal values of precision and recall [32].

$$F1\ Score = 2 * (Precision * Recall)/(Precision + Recall) \quad (5)$$

AUC Score. It stands for area under receiver operator characteristics (ROC) curve. ROC curve is a graph plotting true positive versus false positive [32] (Fig. 3).

4.4 Result Set Statistical Analysis of the Female Kidnapping Records in India

The different purposes for kidnapping and their corresponding age categories and number of female kidnapping for each age category are presented in the following Table 4.

Tabular Analysis:
Table 4 clearly indicates that most of the females below the age of 10 are kidnapped for begging purposes and no women above age of 50 is ever kidnapped for beg-

Table 4 Statistical table for female kidnapping records

Begging		Marriage		Illicit Intercourse		Prostitution		Ransom	
Age	Number of females	Age	Number of females	Age	Number of Females	Age	Number of Females	Age	Number of females
Below 10	256	Below 10	149	Below 10	124	Below 10	13	Below 10	86
10–15	28	10–15	5696	10–15	179	10–15	228	10–15	60
15–18	6	15–18	23,418	15–18	44,046	15–18	826	15–18	64
18–30	36	18–30	75,877	18–30	13,848	18–30	7182	18–30	817
30–50	8	30–50	11,465	30–50	2994	30–50	300	30–50	207
Above 50	0	Above 50	127	Above 50	26	Above 50	51	Above 50	15

Table 5 Measurement of scores for model evaluation

Classes	Accuracy	Precision	Recall	F1 Score	AUC score
Begging	1.0	1.0	1.0	1.0	1.0
Marriage	0.86	1.0	0.33	0.50	0.66
Illicit intercourse	0.93	1.0	0.66	0.80	0.83
Prostitution	0.93	1.0	0.66	0.80	0.83
Ransom	1.0	1.0	1.0	1.0	1.0

ging. Women within age 18–30 are targeted majorly for marriage, and 15–18 are targeted for illicit intercourse. Women with 18–30 are mostly trafficked for prostitution. Statistics for ransom should not possess any relation with age, but still the girls of age 18–30 are mostly targeted. This may be due to the reason that if ransom is not paid then they can traffic those females for prostitution.

Graphical Analysis:
Figures 4a–e represent the graphical charts of Table 4. The charts clearly show that females of age below 10 have more chances for kidnapped for begging, females of age 18–30 are prone for being kidnapped for marriage and prostitution, and females of age 15–18 are targeted for illicit intercourse and ransom. Figure 4f shows the total number of women kidnapped for each of the purposes (begging, marriage, illicit intercourse, prostitution and ransom).

Design of a Predictive Model for Female Kidnapping. The predictive model designed by the application of decision tree using ID3 algorithm is trained properly and subjected to the testing data set. The performance of the evaluation model has to be measured by the use of the testing data set by counting the number of correct classifications and misclassifications. There are certain measures to evaluate a classification model. These measures actually deal with the number of correct classifications and misclassifications and are calculated as ratios. The efficiency of the predictive model has been presented with the following measures:

Tabular Analysis:
Table 5 clearly shows that the kidnappings for begging are accurately predicted, with no false positives and false negatives, with an ideal F1 score and an ideal AUC score. Kidnappings for marriage are predicted with 86% accuracy; prediction has no false positives but large false negatives giving an average F1 score and AUC score. Kidnappings for illicit intercourse are predicted with 93% accuracy, no false positives, moderate false negatives giving a satisfactory F1 score and AUC score. Kidnappings for prostitution and ransom have similar statistics with illicit intercourse and begging, respectively.

Final Average Scores:

- Average accuracy score = 0.94
- Average precision score = 1.0

Graphical Analysis:

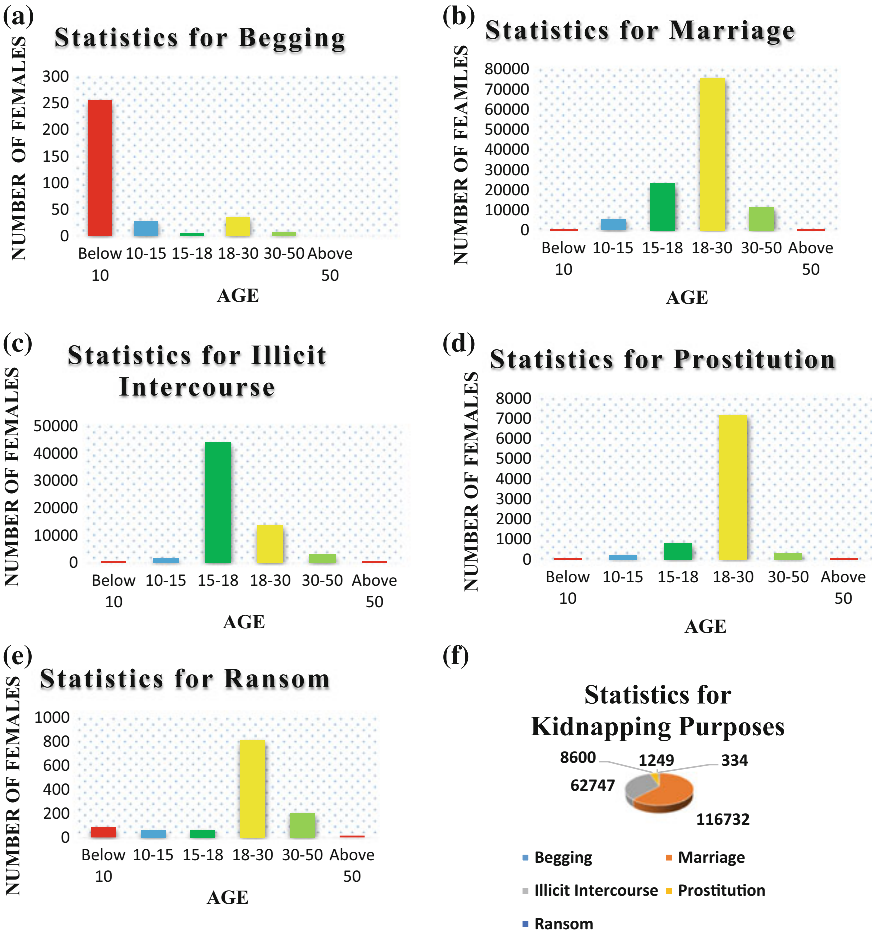


Fig. 4 **a** Begging. **b** Marriage. **c** Illicit intercourse. **d** Prostitution. **e** Ransom. **f** Total female kidnapping

- Average recall score = 0.73
- Average F1 score = 0.82
- Average AUC score = 0.86

Graphical Analysis:

The charts mentioned in Fig. 5a–e clearly state that kidnappings for begging and ransom are predicted with 100% accuracy, without any false positives and false negatives resulting in an ideal F1 score and AUC score. Kidnappings for marriage and illicit intercourse are predicted with 86% accuracy and 93% accuracy, respectively. There are large false negative predictions in cases of marriage and moderate false

Graphical Analysis:

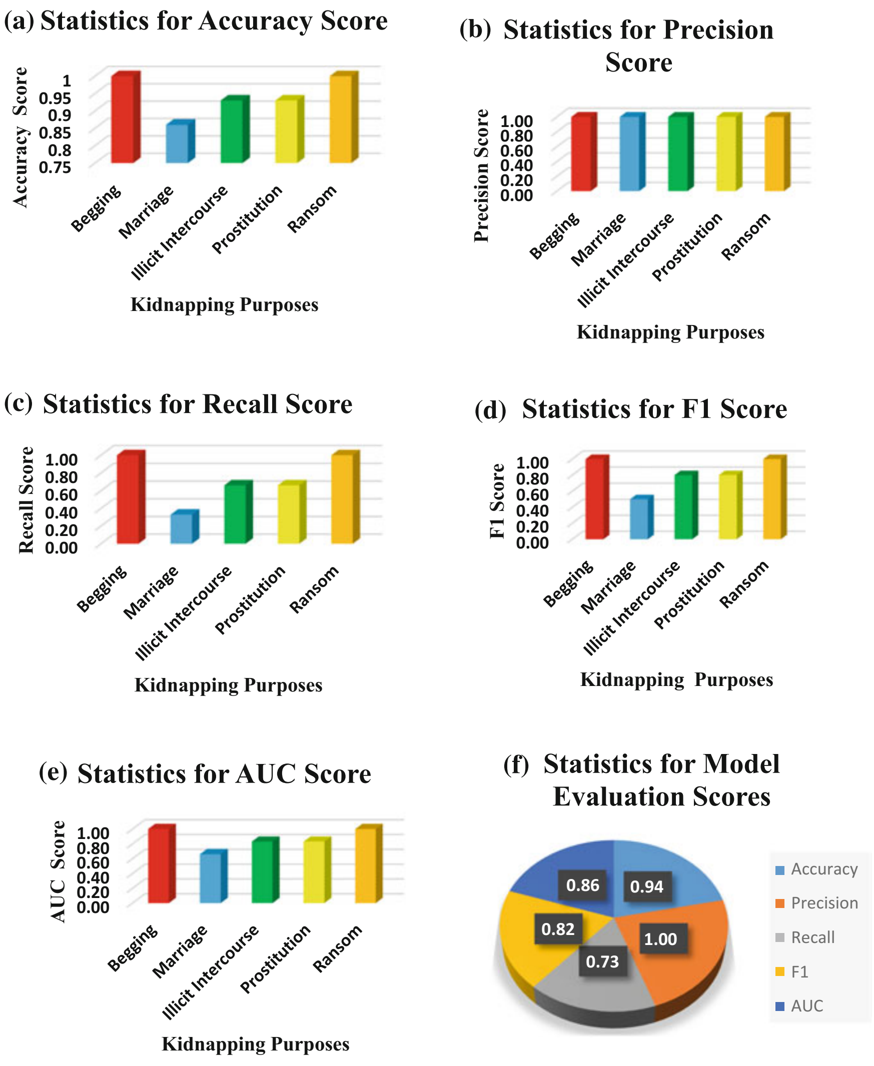


Fig. 5 **a** Accuracy score **b** Precision score **c** Recall score **d** F1 score **e** AUC score **f** Average of all scores

negative predictions for illicit intercourse. F1 and AUC scores for marriage cases are average, while they are satisfactory in cases of illicit intercourse. Statistics for prostitution and ransom are exactly the same as that of illicit intercourse and begging, respectively. Figure 5f gives a description of the average measures of accuracy, precision, recall, F1 and AUC scores after combining all the cases.

5 Conclusion and Future Work

The proposed work after continuous analysis is found to give quite acceptable measures of the parameters of evaluation of the classification model, i.e., accuracy, precision, recall, F1and ROC measures. Thus, it is clear that the proposed predictive model is good enough for the classification and prediction of purposes for kidnapping given the values of certain parameters like age, financial status, glamour and marital status. This model can become an efficient tool for the investigation departments of India to have a rough idea about the purpose and intentions behind the female kidnapping resulting in a better solution of criminal activities related to abduction of women.

However, in order to rely completely on the proposed model, it has to be sure that the model is robust and fault tolerant. The features used as classifiers in the proposed model may not be sufficient enough for generating a very robust and fault-tolerant model. There is a scope of research for searching new or modified features set. The features set used can also be optimized to generate better result. An advanced ensemble-based nature-inspired algorithm can be used in future to generate a new predictive model and compare the result with the proposed model. The robustness of the model can be tested by applying the model on larger and rich data sets and identifying whether there is any abnormal behaviour in the outcome of the model. The fault tolerance can be measured by introducing outliers in the data set and clearly observing the outcomes to find whether there is an increase in the number of misclassifications and change in measures of accuracy, precision, recall, etc. The target is to achieve minimum number of misclassifications. A robust and a fault-tolerant predictive model can be applied in real-life applications.

References

1. Dubey, S., Agarwal, P: Crime, Crime Rates and Control Techniques: A Statistical Analysis
2. Crime Statistic Report of National Crime Record Bureau (N.C.R.B) (2015)
3. Biddle, P., England, P., Peinado, M., Willman, B.: The Darknet and the Future of Content Protection. In: Feigenbaum, J. (eds) Digital Rights Management. DRM 2002. Lecture Notes in Computer Science, vol. 2696. Springer, Berlin, Heidelberg (2003). https://doi.org/10.1007/978-3-540-44993-5_10
4. Status of children in 14–18 years: An Report from National Commission for Protection of Child Rights—Government of India (Published Year 2014 and 2015)
5. NCPR. http://ncpcr.gov.in/showfile.php?lang=1&level=2&&sublinkid=300&lid=739. Accessed 27 Feb 2018
6. McClendon, L., Meghanathan, N.: Using Machine Learning Algorithms to Analyze Crime Data. Mach. Learn. Appl. Int. J. **2**, 1–12. https://doi.org/10.5121/mlaij.2015.2101
7. Burbano, D., Hernandez-Alvarez, M.: Identifying human trafficking patterns online, 2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM), Salinas, 2017, pp. 1–6. https://doi.org/10.1109/etcm.2017.8247461
8. Feinerer, I.: Introduction to the TM Package Text Mining in R
9. Lopez, M., Kalita, J.: Deep Learning Applied to NLP. https://arxiv.org/abs/1703.03091
10. Alvari, H., Shakarian, P., Snyder, J.K.: A Non-parametric Learning Approach to Identify Online Human Trafficking, pp. 33–138 (2016). https://doi.org/10.1109/isi.2016.7745456

11. Dank, M.: Estimating the Size And Structure of the Underground Commercial Sex Economy in Eight Major US Cities (2014)
12. Backpage Homepage (2018). https://www.backpage.com. Accessed 27 Feb 2018
13. Giacobe, N.A., et al.: Characterizing sex trafficking in Pennsylvania for law enforcement. In: 2016 IEEE Symposium on Technologies for Homeland Security (HST), Waltham, MA, pp. 1–5 (2016). https://doi.org/10.1109/ths.2016.7568914
14. Brewster, B., Ingle, T., Rankin, G.: Crawling open-source data for indicators of human trafficking. In: 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, London, pp. 714–719. https://doi.org/10.1109/ucc.2014.116, 2014
15. Boisot, M.: Knowledge Assets: Securing Competitive Advantage in the Information Economy (1998), ISBN-13: 978-0198296072, ISBN-10: 019829607X
16. Ibanez, M., Suthers, D.: Detection of domestic human trafficking indicators and movement trends using content available on open internet sources. In: 2014 47th Hawaii International Conference on System Sciences, Waikoloa, HI, pp. 1556–1565 (2014). https://doi.org/10.1109/hicss.2014.200
17. Kreyling, S., West, C., Olson, J.: Technology and Research Requirements for Combating Human, Pacific Northwest National Laboratory
18. Chhachhiya, V.: Spatio-temporal analysis of kidnapping and abduction of women in Chandigarh. Int. J. Sci. Res. Publ. **7**(7), 648, ISSN 2250-3153 (2017)
19. Rengaraj, V., Bijlani, K.: A study and implementation of smart ID card with M-learning and child security. In: 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, pp. 305–311 (2016). https://doi.org/10.1109/ICATCCT.2016.7912013
20. Huang, Z., Gao, Z., Lu H., Zhang, J., Feng, Z., Xia, H.: An mobile safety monitoring system for children. In: 2014 10th International Conference on Mobile Ad-hoc and Sensor Networks (MSN), pp. 323–328 (2014). https://ieeecomputersociety.org/10.1109/MSN.2014.55
21. Deenadayalan, C., Murali, M., Baanupriya, L.R.: Implementing prototype model for school security system (SSS) using RFID. In: 2012 Third International Conference on Computing Communication & Networking Technologies (ICCCNT), pp. 1–6 (2012). https://doi.org/10.1109/icccnt.2012.6396090
22. Ibanez, M., Gazan, R.:Virtual indicators of sex trafficking to identify potential victims in online advertisements. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, pp. 818–824 (2016). https://doi.org/10.1109/asonam.2016.7752332
23. Kejriwal, M., Szekely, P.: Knowledge graphs for social good: an entity-centric search engine for the human trafficking domain. IEEE Trans. Big Data. https://doi.org/10.1109/tbdata.2017.2763164
24. Cristina, P., Napoca, C.: Decision Tree (2010)
25. CISE, https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm. Accessed 27 Feb 2018
26. Packtub. https://www.packtpub.com/mapt/book/big_data_and_business_intelligence/9781784395803/6/ch06lvl1sec43/entropy-and-information-gain. Accessed 27 Feb 2018
27. The grimmscientist. https://www.thegrimmscientist.com/tutorial-decision-trees/. Accessed 27 Feb 2018
28. Open Government Data: https://data.gov.in/catalog/age-group-wise-victims-kidnapping-and-abduction. Accessed 27 Feb 2018
29. Cran Homepage. https://cran.r-project.org/. Accessed 27 Feb 2018
30. Machine Learning Mastery. https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/. Accessed 27 Feb 2018
31. Saedsayad. http://www.saedsayad.com/decision_tree.htm. Accessed 27 Feb 2018
32. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/. Accessed 27 Feb 2018
33. Codalab. https://codalab.lri.fr/competitions/132#learn_the_details-evaluation. Accessed 27 Feb 2018