Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020)
IEEE Xplore Part Number:CFP20K25-ART; ISBN:978-1-7281-4889-2

Diagnosis of Crime Rate against Women using k-fold Cross Validation through Machine Learning Algorithms

P. Tamilarasi M.C.A., M.Phil
Ph.D Research Scholar,
Department Of Computer Science,
Sri Sarada College For Women (Autonomous),
Salem -16.
tamilinresearch@Gmail.Com

Dr.R.Uma Rani M.C.A., M.Phil., Ph.D
Principal,
Sri Sarada College For Women (Autonomous),
Salem -16.
umainweb@gmail.com

***Abstract*— Crime against women has become a very big problem of our nation. Many countries are trying to control this offence continuously and its prevention is an essential task. In recent years crimes are significantly increasing against women. Currently the Indian government show interest to address this problem and give more importance to develop our society. Every year a huge amount of data collection is generated on the basis of the crime reporting. This data can be very useful for assessing and predicting crime, and can help us to some degree stop the crime. Data analysis is a process of examining, cleansing, transformation and modelling data with the goal of establish useful information, reporting conclusion and sustaining decision-making. Feature Scaling is one of the most important techniques to standardize the independent features to place the data in a fixed range. It is performed at the time of data pre-processing. K-fold cross-validation is a re-sampling method used for calculating machine learning models on a small sample of data. It is a common strategy since it is easy to understand and usually results in a model deftness calculation that is less biased or less negative than other approaches, such as a simple train or test divide. Machine learning plays a large part in data processing. This paper introduces six different types of Machine learning algorithms such as KNN and decision trees, Naïve Bayes, Linear Regression CART (Classification and Regression Tree) and SVM using similar characteristics on crime data. Those algorithms are tested for accuracy. The main objective of this research is to evaluate the efficacy and application of the machine learning algorithms in data analytics.**

Keywords: Crime Rate, Feature scaling, k-fold cross validation, Machine Learning Algorithms

I. INTRODUCTION

In worldwide Crime level is increasing every day. Offense can't be expected since it is either proficient or unplanned. Crime is a cost-effective trouble distressing life value and economic escalation. The particulars of how crime is performed revolutionize depending on the type of nation and society. Prior researches in crime prediction have originated that factors resembling education, deficiency, employment, and circumstances affect the crime rate. The word aggression used to state huge range of acts. It includes POCSO (protection of child from sexual offences act), DPA (dowry prohibition act), cruelty by husband, dowry death, molestation and TN prohibition of women harassment. In WHO report that about 35% of women distress by sexual violence. Globally 38% of women murders are committed by hers partner. 7% of sexual cruelty occurs by other than a spouse. Newly, the Vancouver Police Department (VPD) instigates a crime analytical model to predict crimes. The techniques of crime prediction law enforcement used to identify doubtless crimes. Globally many researchers have been continued in this area. Machine learning is the scientific and statistical type of algorithms which are recently used in image recognition, speech recognition, medical diagnosis, statistical arbitrage and classification. It's also made for predicting crime rate at a specific year which based on crime against women data. The main aim of this work is to develop prediction model that can be used to predict crime rate accurately. In this work implemented different types of machine learning algorithms used to analyze crime against women data. This data collected from 2001 to 2012 with more than 2, 00,000 records. This research followed three different types of methods to process the data set.

- Data preprocessing is the first step. In this method removed the blank spaces, redundant data and rescaling the data using standardization.
- In the Second step, proposed k-fold cross validation.
- In the third step, predict crime rate and accuracy using different types of ML algorithms.

II. RELATED WORKS

Many researchers have discussed about the problems of crime control and have proposed different algorithms for crime prediction. The accuracy of prediction and classification is based on the attributes in features and the dataset used as a reference. The author H. Chen and W. Chung developed crime based frame work using different data mining techniques. To develop this project the author used three models. These model names are named-entity extraction, deceptive-identity detection, and criminal-network analysis [1]. Prabakaran, S., & Mitra, S. discussed about four types of crime i.e. Fraud detection, traffic violence, violent crime, web crime and

sexual offense using different mining algorithms [2]. The author applied neural networks and clustering techniques for predict crime. By using these techniques, crime data can be automatically stored in the law enforcement agencies database [3]. In this work the author used KNN classification algorithm for predict specific crime region which have high level probability for crime occurrence's [4].In [5], the author proposed clustering and classification techniques to anticipate crime trends and analysed the individuals activities by link analysis .In [6], The author predicts the metropolitan urban areas violations rate to decrease and anticipate through clustering in WEKA tool with Euclidian distance. The authors McClendon, L., & Meghanathan, N implemented the Linear Regression, Additive Regression, and Decision Stump algorithms for finding accuracy on crime data based on correlation coefficient [7]. In [8], the writer performed the comparative study between the different machine learning algorithms to prove that algorithms accuracy and efficiency. In [9], the author executed logistic regression model to explore the relationship between several predicting feature and burglary happening probability with an observation to the epicenter. In [10], the author implemented KNN and decision tree algorithms for predict the accuracy using Vancouver crime data. In [11], the researcher proposed a novel transfer learning method to incorporate spatio-temporal patterns in urban data in a specific area, and then exploit transfer learning techniques to illustrate other region's crime prediction. In [12] this paper author explores models for predicting the regularity of different types of crimes by LSOA code (Lower Layer Super Production Areas — an administrative network of areas used by the UK police) and the incidence of anti-social tag crime. In this work three separate algorithm types are trees of applied learning, regression and decision.

III. PROPOSED METHODOLOGY

A. Data Preprocessing

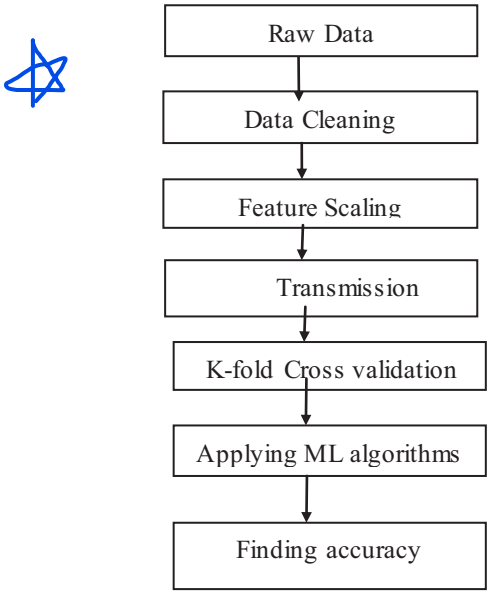

Fig. 1: Work Flow of Crime Prediction

It is a one of the most important techniques. This is used to transform raw data into a clear format. Raw data (real world data) cannot be applied through a model because it is always imperfect and it would cause certain errors. That is why we should preprocess the data before sending to the model.

A.1 Data cleaning

This method removes main errors and contradiction that are expected when multiple resources of data are getting into the dataset, where mean values are used at the place of NAN in the required attribute.

A.2 Feature Scaling the Data

When our data consists of different scales of attributes, many machine learning algorithms can benefit from rescaling the attribute to all have similar scale. This provides us values between 0 and 1 standardization. The below given formula is used to standardize the variables.

$$z = \frac{X - \mu}{\sigma} \quad (1)$$

Steps for finding Standard deviation
Step1: Subtract the mean value from the value X. X is an observations.
Step 2: Divide the result values from Step1 by σ.

Data transformation is the method of converting data from one format into another format. Data transformation is significant to tricks such as data integration and data management. PCA is statistical techniques. It is used to identify the uncorrelated data from larger data set.

Cross-validation is statistical. It is primarily used for selecting the model to estimate a predictive model's test error better. The principle of cross validation is to split the sample observation as number of groups. It is a common approach since it is easy to understand and leads to a less biased or less positive assessment of the ability of the model than other methods. Various types of cross validations are available. This paper uses k-fold cross validation techniques to validate the results.

In k-fold cross validation techniques all the training data set are considered for both training and validation. In training data set all entries are used for validation.

Steps for k-fold cross validation:

Step 1: Divide the training data set into k equal subsets like f1, f2, f3….fk. Here all subset is called a fold.
Step 2: For i=1 to i=k
Step 3: Consider f as validation set and all the remaining set K-1 folds are in the cross validation training set.
Step 4: Using cross validation train the ML model and calculate the accuracy
Step 5: Evaluate the accuracy using all the k cases of cross validation.

This way involves randomly separating the set of observations into k groups or folds, of roughly equal size. The first fold is assign as a validation set, and the remaining k − 1 fold fit on the method. It is also essential that any instruction of the data earlier to fitting the model occur on the CV consign training dataset within the loop rather than on the large data set.

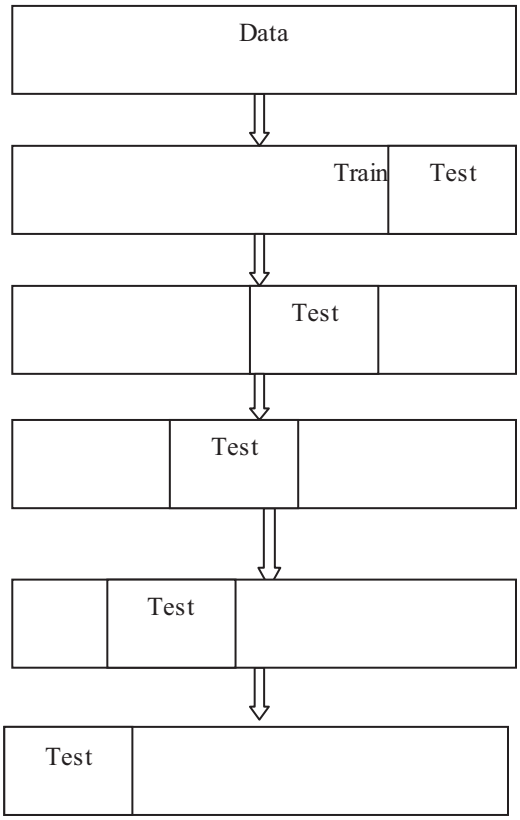

Fig. 2: Diagram for k-fold 5 cross validation

This is also applied to any modification of hyper parameters. A failure to make this process within the loop may result in data leakage and an optimistic evaluation of the model skill. Fold cross validation results are often re-iterated with mean of the model values. It is also good performance to consist of a measure of the variance to the values, such as the standard error or standard deviation. The below given formula gives the cross validation error.

$$CV(\lambda) = \frac{1}{K} E_k(\lambda) \quad (2)$$

When the model gives minimum mean squared error on the training data it can be optimistic to predictive error.

B. Machine Learning Model.

A model of Machine Learning is a mathematical depiction of a real world. The result of the training process is a machine learning model which is used to make predictions. In this paper various types of ML algorithms such as KNN, NaiveBayes, Linear Regression ,classification and Regression Tree(CART),Support Vector Machine, K Nearest Neighbour(KNN),Linear Discriminate Analysis(LDA) using Python to predict the accuracy for crime data.

IV. RESULT AND DISCUSSIONS

The below picture clearly explained about highest Crime Rate of given data from 2001 to 2012.This crime against women records are collected from Data.Gov.in.It has different types of crimes with 29 Indian states

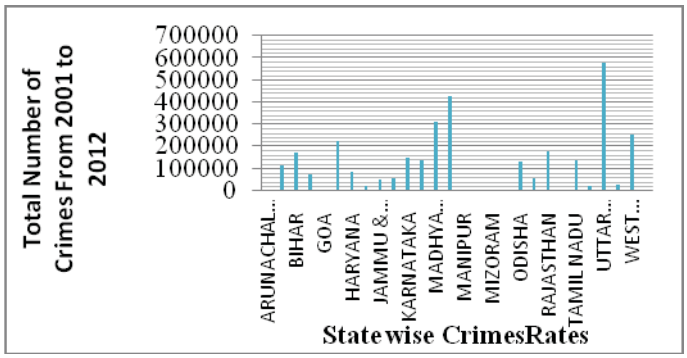

Fig. 3: Indian State wise Crime details

The above fig.3 shows the maximum value of crimes happened against women in utterpradesh,

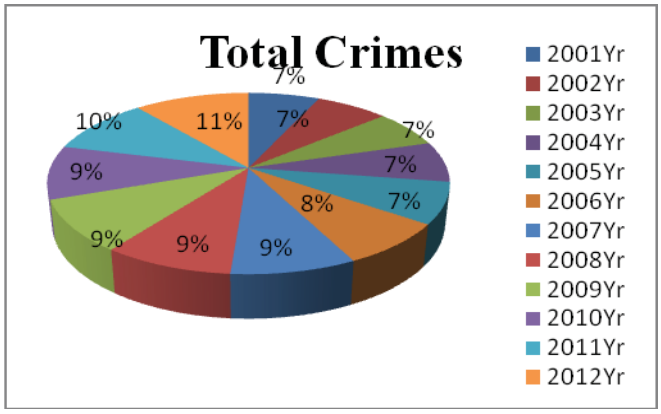

Fig. 4: Year wise Crime details

India In 2012, the highest number of crimes happened against women. Fig.4 shows this detail clearly

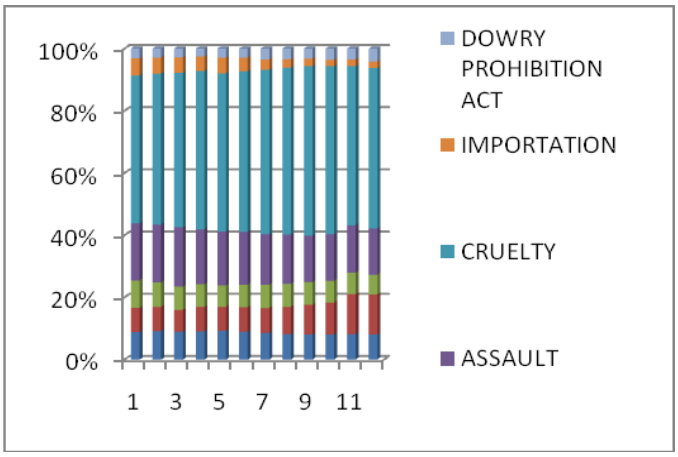

Fig.5: Highest Crime details

The above picture expose cruelty crime type happened maximum in the period of 2001 to 2012.

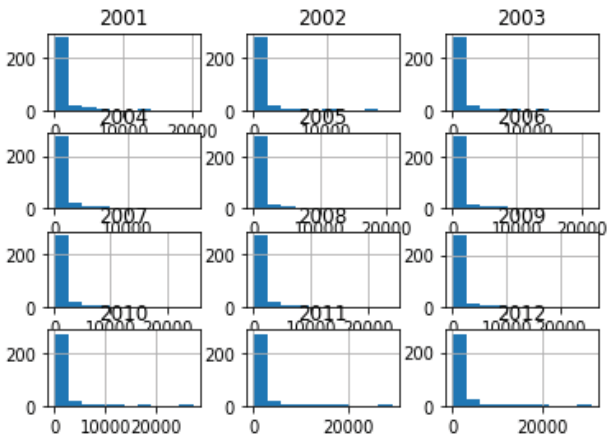

Fig. 6: Histogram plot for Crime Data

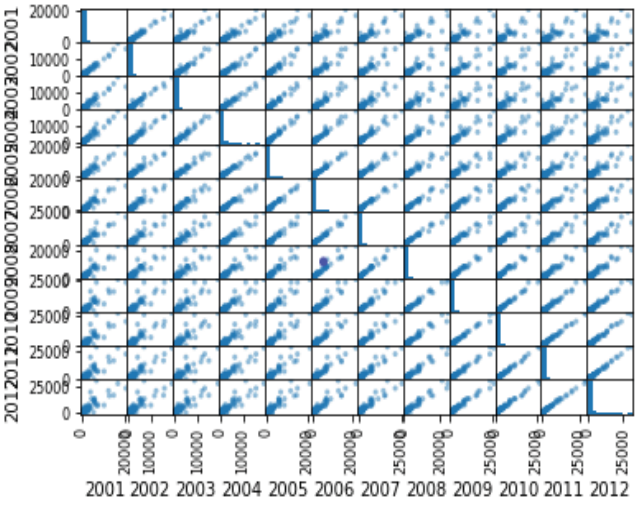

Fig. 7: Multivariate plot for Crime Data

The above fig.6 and 7 are represent how the violence frequently increasing against women since 2001 to 2012.

Table I.

Types of ML Algorithm	MEAN	Standard Deviation
NB	0.284045	0.071719
CART	0.260504	0.07998
KNN	0.300385	0.074799
LDA	0.230672	0.08339
LR	0.26348	0.054434
SVM	0.218639	0.050919

Machine learning algorithms Performance

The above Table.I shows the different types of ML algorithm performance.

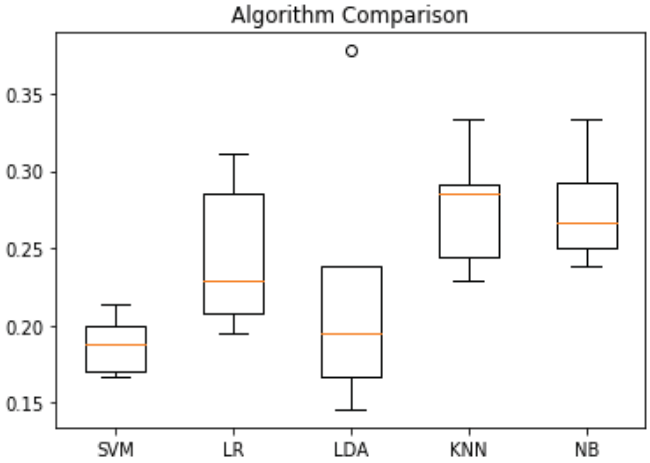

Fig.8: ML algorithm performance comparison using box plot

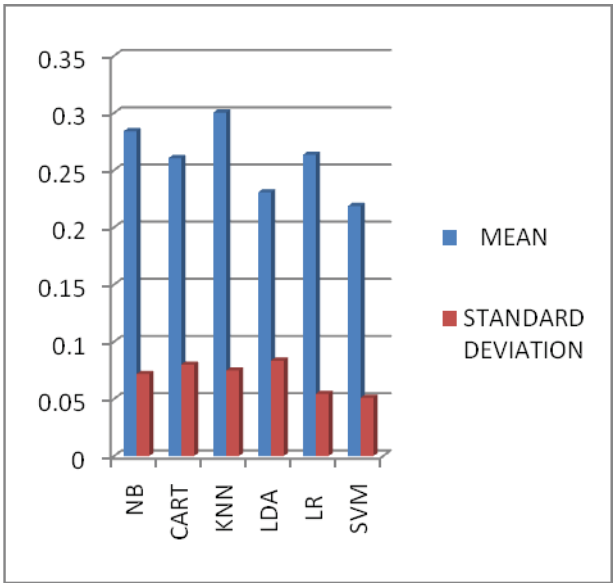

Fig. 9: ML algorithm efficiency comparison Based on Mean and Standard Deviation

From this above result the Knn is performed better than other ML algorithms. The above fig. 8 and 9 shows the machine learning algorithms accuracy in graphical representation.

V.CONCLUSION

From these above results it concluded that the algorithm of KNN is efficiency than other ML algorithms. This paper expressed highest crime region and which crime type is happened regularly in India. This result will be helpful to control the violence against women in future. Further we can continue this work using different types of cross validation techniques for better accuracy.

REFERENCES

[1] Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., & Chau, M. (2004). *Crime data mining: a general framework and some examples. Computer, 37(4), 50–56.* doi:10.1109/mc.2004.1297301

[2] Prabakaran, S., & Mitra, S. (2018). *Survey of Analysis of Crime Detection Techniques Using Data Mining and Machine Learning. Journal of Physics: Conference Series, 1000, 012046.*doi:10.1088/1742-6596/1000/1/012046

[3] Keyvanpour, M. R., Javideh, M., & Ebrahimi, M. R. (2011). *Detecting and investigating crime by means of data mining: a general crime matching framework. Procedia Computer Science, 3, 872–880.*doi:10.1016/j.procs.2010.12.143

[4] Sathyadevan, S., S, D. M., & S., S. G. (2014). *Crime analysis and prediction using data mining. 2014 First International Conference on Networks & Soft Computing (ICNSC2014).*doi:10.1109/cnsc.2014.6906719

[5] Yamuna's, N.Sudha Bhauvaneswari D, Data mining Techniques to Analyze and Predict Crimes, International Journal of Engineering And Science(IJES) Vol-1,Issue-2,PP 243-247

[6] S. Lavanyaa, D. Akila . Crime against Women (CAW) Analysis and Prediction in Tamilnadu Police Using Data Mining Techniques , International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-5C, February 2019

[7] Alves, Luiz GA, Haroldo V. Ribeiro, and Francisco A. Rodrigues. "Crime prediction through urban metrics and statistical learning." *Physica A: Statistical Mechanics and its Applications* 505 (2018): 435-443.

[8] Mol, P. Roshni, and C. Immaculate Mary. "Intrusion Detection System from Machine Learning Perspective."\

[9] Alaoui, Safae Sossi, Brahim Aksasse, and Yousef Farhaoui. "Data Mining and Machine Learning Approaches and Technologies for Diagnosing Diabetes in Women." In *International Conference on Big Data and Networks Technologies*, pp. 59-72. Springer, Cham, 2019.

[10] Duraipandian, M. "Performance Evaluation of Routing Algorithm for MANET based on the Machine Learning Techniques." *Journal of trends in Computer Science and Smart technology (TCSST)* 1, no. 01 (2019): 25-38.

[11] Galán-García, Patxi, José Gaviria de la Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying." *Logic Journal of the IGPL* 24, no. 1 (2016): 42-53.

[12] Sessink, Danique. "Using Machine Learning to Detect ICT in Criminal Court Cases." Bachelor's thesis, University of Twente, 2018.