Proceedings of the 2nd International Conference on Trends in Electronics and Informatics (ICOEI 2018)
IEEE Conference Record: # 42666; IEEE Xplore ISBN:978-1-5386-3570-4

Crime Data Analysis and Prediction of Perpetrator Identity using Machine Learning Approach

Mary Shermila A
Loyola-ICAM College of Engineering & Technology
Loyola Campus, Nungambakkam,
Chennai, Tamil Nadu, India.
maryshermila.18cs@licet.ac.in

Amrith Basil Bellarmine
Loyola-ICAM College of Engineering & Technology
Loyola Campus, Nungambakkam,
Chennai, Tamil Nadu, India.
amrithbasil.18cs@licet.ac.in

Nirmala Santiago
Loyola-ICAM College of Engineering & Technology
Loyola Campus, Nungambakkam,
Chennai, Tamil Nadu, India.
nirmala@licet.ac.in

***Abstract*—Crime is one of the most predominant and alarming aspects in our society and its prevention is a vital task. Crime analysis is a systematic way of detecting and investigating patterns and trends in crime. The aim of this model is to increase the efficiency of crime investigation systems. This model detects crime patterns from inferences collected from the crime scene and predicts the description of the perpetrator who is likely suspected to commit the crime. This work has two major aspects: Crime Analysis and Prediction of perpetrator identity. The Crime Analysis phase identifies the number of unsolved crimes, and analyses the influence of various factors like year, month, and weapon on the unsolved crimes. The prediction phase estimates the description of the perpetrators like, their age, sex and relationship with the victim. These predictions are done based on the evidences collected from the crime scene. The system predicts the description of the perpetrator using algorithms like, Multilinear Regression, K-Neighbors Classifier and Neural Networks. It was trained and tested using San Francisco Homicide dataset (1981-2014) and implemented using python.**

Keywords— Multilinear Regression; K-Neighbors Classifier; Artificial Neural Networks.

I. INTRODUCTION

A crime is an action which constitutes an offence and is punishable by law. Analyzing and Identifying the hidden crime patterns are the major problems to the police department as there are voluminous data of crime exist. So, we need some methodologies to help the investigation bureau in solving the crimes.

Machine Learning approach provides regression and classification techniques that helps us to serve the purpose. Regression techniques like Multi-linear regression, is a statistical method which is used to find the relationship between two quantitative variables. This predicts the value of the dependent variables(y) based on the independent variables(x). Classification techniques like K-Neighbors Classification is used to classify the multiclass target variables. Neural networks are then used to improve the accuracy of the predictions, which has an input layer, dense layer and an output layer. Based on these algorithms the perpetrator description like age, sex and relationship with the victim is predicted by the model. This system is thus expected to ease the burden of the police department in solving the homicide cases.

II. RELATED WORKS

Ling Chen, Xu Lai (2011) [1] compared the experimental results obtained by Artificial neural Network (ANN) and Autoregressive Integrated Moving Average (ARIMA) in forecasting the hourly wind speed. On Comparison, ANN model produces a better result when compared to ARIMA model.

Jyoti Agarwal, Renuka Nagpal et al., (2013) [2] has done Crime analysis using k-means clustering on the crime dataset. This model is developed using rapid miner tool. The clustered results are analyzed by plotting the values over the years. The model thus concludes from the analysis that the number of homicides decreases from 1990 to 2011.

Shiju Sathyadevan, Devan M. S et al., (2014) [3] predicted the regions which have high probability for crime occurrence and visualized crime prone areas. The authors classified the data using the Naive Bayes classifiers algorithm which is a supervised learning as well as a statistical method for classification and has provided 90% accuracy.

Lawrence McClendon and Natarajan Meghanathan (2015) [4] used various prediction algorithms such as Linear Regression, Additive Regression, and Decision Stump algorithms using the same set of input (features), on the Communities and Crime Dataset. Overall, the linear regression algorithm gave the best results compared to the three selected algorithms. The main advantage of linear regression algorithm is, it could handle randomness in the test data to a certain extent (without incurring too much 15 of prediction error).

Rasoul Kiani, Siamak Mahdavi et al., (2015) [5] proposed a framework for predicting the crimes by using clustering algorithms. This is implemented using RapidMiner tool. In order to increase the efficiency of prediction, GA (Genetic Algorithm) is used for detecting the

outliers in the data. This model has produced an accuracy of 91.64%.

Ryan Heart field, George Loukas et al., (2016) [6] predicts the rate of crimes that occurs due to Semantic Social Engineering Attacks and explores the feasibility of predicting user susceptibility to deception-based attacks. The authors have predicted using logistic regression and a random forest prediction model, with the accuracy rates of .68 and .71, respectively.

S. Sivaranjani, S. Sivakumari et al., (2016) [7] used various clustering approaches like the K-Means clustering, Agglomerative clustering and Density Based Spatial Clustering with Noise (DBSCAN) algorithms are used to cluster crime activities in Tamil Nadu. The performance of each clustering algorithms is evaluated using the metrics such as precision, recall and F-measure, and the results are compared. Based on the above metrics, DBSCAN algorithm gave the best results compared to the other two selected algorithms.

Chirag Kansara, Rakhi Gupta et al., (2016) [8] proposed a model which analyze the sentiments of the people in twitter and predicts whether they can become threat to particular person or society. This model is implemented using Naive Bayes Classifier which classifies the people by sentiment analysis.

III. METHODOLOGY

A. Dataset

The crime dataset is obtained from Kaggle which a repository of datasets on various domains is. This dataset consists of homicide entries collected from

(i) The FBI's Supplementary Homicide Report from 1980 to 2014 and

(ii) Right to Information Act data on more than 22,000 homicide entries that were not solved by the Investigation department. The dataset consists of 638454 rows and 17 columns and the column metadata is given in TABLE I.

From the dataset, the significant features like State, Year, Month, Crime Type, Crime Solved, Victim Gender, Victim Age, Victim Race, Victim Count and Weapon are chosen as the input features for the system

The features Perpetrator Age, Perpetrator Sex and Relationship of the perpetrator with the victim are chosen as the target variable to be predicted by the system.

TABLE I COLUMN METADATA

No	Name	Type Of Column	Description
1	Record ID	Numeric	Unique Id of the case.
2	Agency Code	String	The code of the Agency which handled the case.
3	Agency Name	String	The name of the Agency which handled the case.
4	Agency Type	String	The type of the agency say municipal police, sheriff etc.
5	State	String	The state of occurrence of the crime and it has around 50 unique states.
6	Year	Numeric	The year in which the crime has occurred.
7	Month	String	The month in which the crime has occurred.
8	Crime Type	String	The type of the crime
9	Crime Solved	String	Status of the investigation
10	Victim Sex	String	The gender of the victim
11	Victim Age	Numeric	Age of the victim
12	Victim Race	String	The Race of the victim
13	Weapon	String	Weapon used for committing the crime
14	Victim Count	Numeric	No of victims in the case
15	Perpetrator Age	Numeric	Age of the perpetrator
16	Perpetrator Sex	String	Gender of the perpetrator
17	Relationship	String	Relationship of the perpetrator with the victim

B. WorkFlow Diagram

Fig. 1 explains the workflow of the system. The workflow starts by extracting the homicide data from Kaggle, which is a repository of datasets on various domains. The raw data is then preprocessed and converted into a crime database.

The database is then provided to crime analysis phase and prediction phase. The crime analysis helps in analyzing

the unsolved crimes in the database. This is done using python library called Matplotlib which visualizes the data on various aspects given to it.

The prediction is done for three target variables.

- The perpetrator age is predicted using Multi Linear Regression.
- The perpetrator sex is predicted using K-Neighbors Classifier and Neural Networks.
- Relationship is predicted using K-Neighbors Classifier and Neural Networks. The results of predictions are discussed, and accuracy is tested.

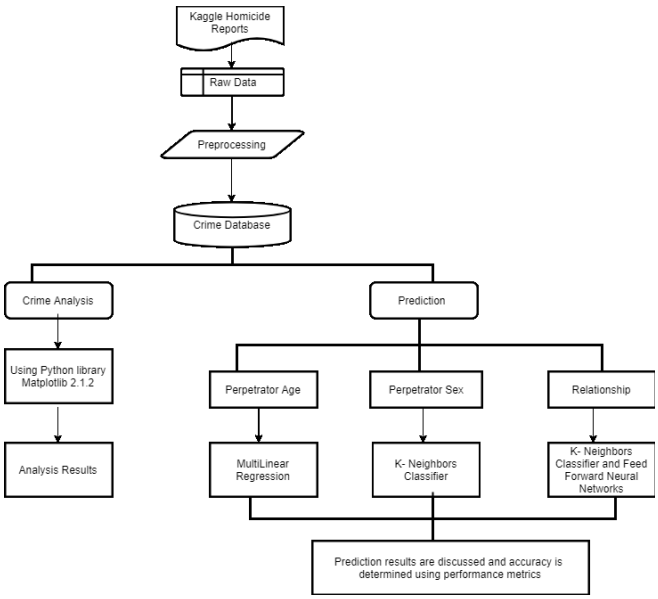

Fig. 1. WorkFlow Diagram

C. Preprocessing

In the above dataset some of the features like Month, Crime Type, Crime Solved, Victim sex, Victim Age, Victim Race, Weapon, Perpetrator Age, Perpetrator Sex and Relationship are in qualitative form. This qualitative data should be categorized as 0s and 1s in order to apply the mathematical models for prediction. Preprocessing is done by adding dummy columns [9]. This method adds N-1 dummy columns for N unique values in the given column.

TABLE II PREPROCESSING

States	Dummy_A	Dummy_B
Alaska	1	0
Alabama	0	1
Anchorage	0	0

The above table explains how the preprocessing of data is done. Three states namely Alaska, Alabama and Anchorage from the Column State in the dataset are chosen. In this case the number of unique values in the column state is N=3. To categorize the data N-1 dummy columns should be added. Therefore, the number of dummy columns is 2, which is Dummy_A and Dummy_B.

- If the state is Alaska, then assign 1 for Dummy_A and 0 for Dummy_B.
- If the state is Alabama, then assign 0 for Dummy_A and 1 for Dummy_B.
- If the state is Anchorage, then assign 0 for both Dummy_A and Dummy_B.

This is implemented with the help of LabelEncoder and OneHotEncoder in python. This creates N dummy columns from which one column can be dropped randomly to attain N-1 dummy columns. This is verified using Trial and Error method in which columns are chosen randomly and dropped in iterations. The model produces the same accuracy value in each trial.

IV. IMPLEMENTATION

A. Analysis

The dataset that we had taken into consists of 638454 crime entries between the years 1980 and 2014. The Analysis phase analyzes and identifies,

- The number of unsolved crimes
- Weapons used in the unsolved crimes
- The month in which the maximum number of unsolved crimes have occurred, and
- The investigation body has more number of unsolved crimes

It is observed that out of 638454 crime entries, there were 190282 unsolved crimes reported across the various states and counties entailed within the dataset.

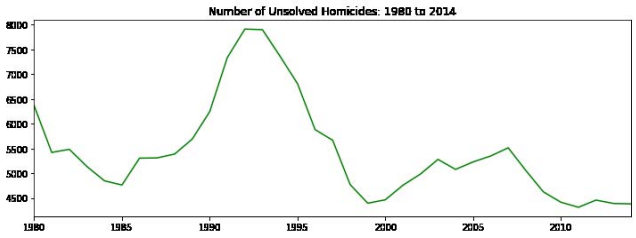

Fig. 2. Years vs. No of Unsolved Crimes

Fig. 2 shows a sharp increase in the number of unsolved crimes between 1990 and 1995 reaching an all-time high in the year 1993.

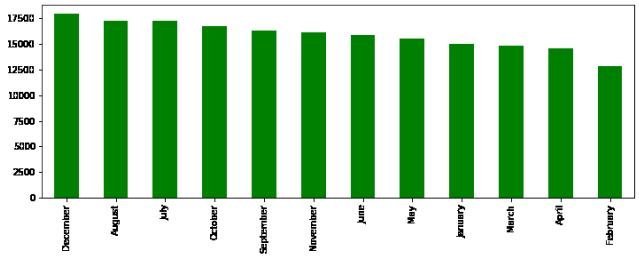

Fig. 3. Month vs.No of Unsolved Crimes

Similarly, Fig. 3 shows that December month stands first with 1, 75, 000 unsolved crimes, followed by August and July with comparatively equal no of unsolved crimes.

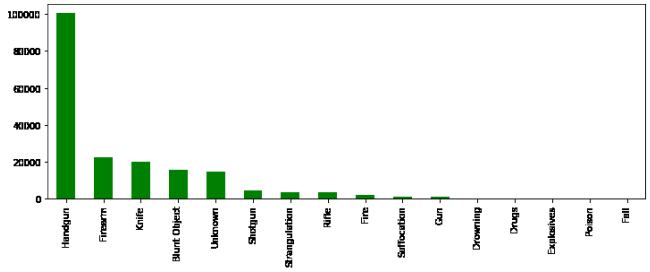

Fig. 4. Weapons vs. No of Unsolved Crimes

As seen in Fig. 4 the maximum no of unsolved crimes (approx. 1, 00,000) was committed with the help of handgun (weapon). This shows that handgun is predominantly available for committing the crimes in San Francisco.

From the analysis, it is also derived that the agency which had reported the most number of unsolved crimes was the Municipal Police as seen in Fig. 5

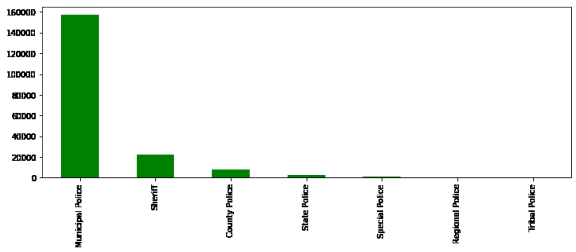

Fig. 5. Investigation Bureau vs. Unsolved Crimes

Thus, the prediction phase of this work aims to find the plausible description of the most likely perpetrator. This description could be used by crime investigation agencies or prosecutorial agencies to help build a case or aid in providing leads to unsolved crimes.

B. Algorithms

a) MultiLinear Regression

Multilinear regression is a mathematical approach for finding relationship between a dependent variable (perpetrator age) with a given set of independent variables (input evidences collected from the crime scene) [10]. This method predicts the value of perpetrator age based on the input features which are mentioned in column metadata Fig.1. The equation for the Multilinear Regression line is given as:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \quad \textbf{(1)}$$

Where,

Y is the dependent variable

x is the independent variable

β_i are coefficients of the regression equations

In the scenario of the crime prediction system, linear regression is used to determine the most likely perpetrator age given the information obtained from the crime scene. Subsequently, an r2_score () of 0.60 with linear regression was obtained.

b) K-Neighbors Classification

K-Neighbors classifier is used when the target variable has more than two classes to be classified [11]. In this dataset, the target variable perpetrator sex has three classes namely male, female, and unknown. Similarly, the target variable relationship has 27 unique classes like friend, husband, wife etc. Hence, K-Neighbors Classifier is used to classify these target variables (perpetrator sex and relationship).

Pseudo Code:

K_Nearest_Classifier (input variables);

Assign K -> the number of clusters

A set of K instances are chosen to be centers for the clusters

For each data point in the input:

Calculate the Euclidian distance

Assign the cluster which is near to the data point

Recalculate the centroids and reassign the variables in the clusters

Repeat until a suitable cluster population is reached.
Return the clusters and the values in it.

Euclidian distance between two data points (p_1, q_1) and (p_2, q_2) is calculated as follows:

$$\mathbf{d\ (p, q)} = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad \textbf{(2)}$$

Using the K-Neighbors Classifier,

- the perpetrator sex is predicted with an accuracy of 0.85 and
- the relationship of the perpetrator with the victim is predicted with an accuracy of 0.48

The accuracy of both the attributes are evaluated using the metric accuracy_score ().

c) Artificial Neural Network

To improve the accuracy of crime prediction [12], we had employed the use of artificial neural networks in our system. The Artificial Neural Networks works like the human brain. The input to the neuron is given as

$$\mathbf{Z = w*x + b} \quad \textbf{(3)}$$

Where x is the input, w is the weights assigned for the input and b is the bias value. Concerning the number of neurons in the hidden layer, the rule of thumb states that, it should be

(a) Between the input and output layer size, or

(b) Set to something near (inputs + outputs) * 2/3, or

(c) Never larger than twice the size of the input layer.

TABLE III SPECIFICATIONS OF NEURAL NETWORK

Specification of the Neural Networks	Attributes	
	Perpetrator Sex	Perpetrator Relationship with the Victim
No of Input Units	98	98
No of Output Units	3	27
No of Hidden Layers	1	3
No of Training Steps (epochs)	19	9
Batch Size	100	100
Activation Function	Sigmoid	Sigmoid
Loss	Binary Cross Entropy	Binary Cross Entropy
Performance Metrics	Accuracy_score ()	Accuracy_score ()
Accuracy of Prediction	0.96	0.97

Initially, the weights are chosen randomly and based on the loss; the network back propagates and adjusts the weights accordingly.

With respect to our system, ANN is implemented with the specifications mentioned in Table III. ANN is used for predicting the perpetrator sex and relationship of the perpetrator with the victim.

C. Performance Metrics

1) r2_score ()

The coefficient of determination or r2_score function tells how well the predicted values match with the actual output values in terms of Regression model.
Let N be the Number of Samples, y_i is the actual value and $\hat{y}_i$ is the predicted value. Then,

$$\mathbf{r2\_score} = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y})^2} \quad \textbf{(4)}$$

2) accuracy score()

The accuracy score function is the most predominant performance measure for evaluating the performance of a multilabel classification model and ANN model. This metric returns the subset accuracy. If the entire set of predicted values for a sample strictly matches with the actual values, then the subset accuracy is 1.0; or else based on the similarity, a score will be returned between 0 and 1.

Let N be the Number of Samples, y_i is the actual value and $\hat{y}_i$ is the predicted value. Then,

$$\mathbf{Accuracy\ score} = \frac{1}{N} \sum_{i=0}^{N-1} 1(\hat{y}_i = y_i) \quad \textbf{(5)}$$

3) Precision

The precision is the ratio of True Positives (TP) to True positives (TP) + False Positive (FP).

$$\mathbf{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad \textbf{(6)}$$

4) Recall

The recall is the ratio of True Positives (TP) to True positives (TP) + False Negatives (FN).

$$\textbf{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (7)$$

5) F1_score

$$\textbf{F1\_Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (8)$$

6) Jaccard Similarity Score

The jaccard_similarity_score function calculates the sum of Jaccard similarity coefficients between pairs of predicted values. The Jaccard similarity coefficient of the i-th samples, with actual value y_i and predicted value $\hat{y}_i$, is defined as:

$$\mathbf{J} = \frac{y_i \cap \hat{y}_i}{y_i \cup \hat{y}_i} \quad (9)$$

7) Receiver Operating Characteristics

The roc_auc_score function covers the area below the receiver operating characteristic (ROC) curve. The ROC curve, is a plot which defines the performance of a classification system as its discrimination threshold is varied.

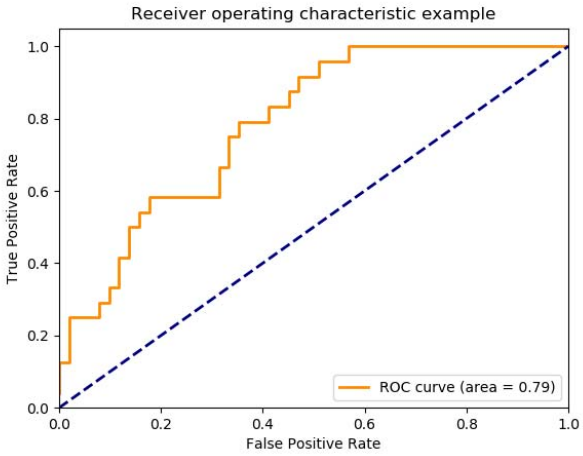

Fig. 6. Representation of ROC Curve

D. Loss Metrics

1) Hamming Loss

The Hamming Loss computes the average hamming distance between the predicted values and the actual values. The Hamming Loss is given by

$$\text{Hamming Loss} = \frac{1}{N}\sum_{i=0}^{N-1}(\hat{y}_i \neq y_i) \quad (10)$$

Where, $\hat{y}_i$ is the predicted output
y_i is the corresponding true value, and
N is the number of samples

2) Zero one Loss

The zero_one_loss function computes the sum or the average of the 0-1 classification loss over the samples N. In multilabel classification, the function returns zero if there are any errors. By default, the function returns the percentage of incorrectly predicted subsets.

$$\text{Zero Loss} = 1\ (\hat{y}_i \neq y_i) \quad (11)$$

Where, actual value is y_i and predicted value is $\hat{y}_i$.

E. Performance Evaluation

1) Performance Analysis of Perpetrator Age using Multilinear Regression

The performance Analysis of a Multilinear Regression model is evaluated using the metric R2_Score.

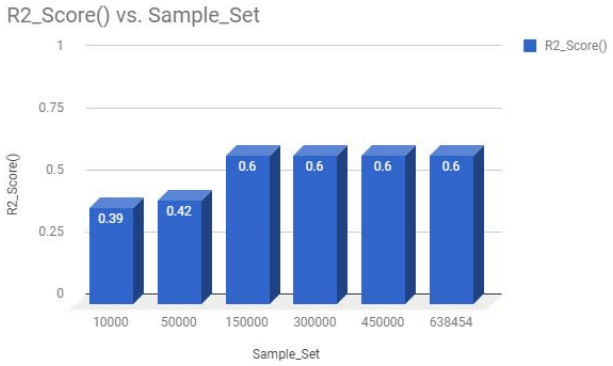

Fig. 7. Prediction results of Perpetrator Age

Initially, It's observed that a sample set of 10000 records bears a meagre r2_score value of 0.39. On increasing the sample set by five times, an r2_score of 0.42 is observed. However, the score reaches a peak of 0.6 when a sample set size of 150000 is reached. The value remains constant even at a sample set size of 665000.

2) Performance Analysis of Perpetrator Gender using K-Neighbors Classifier

The Performance analysis of Perpetrator Gender is done through Precision, Recall, F1_Score, ROC, Accuracy score and Jaccard Score.

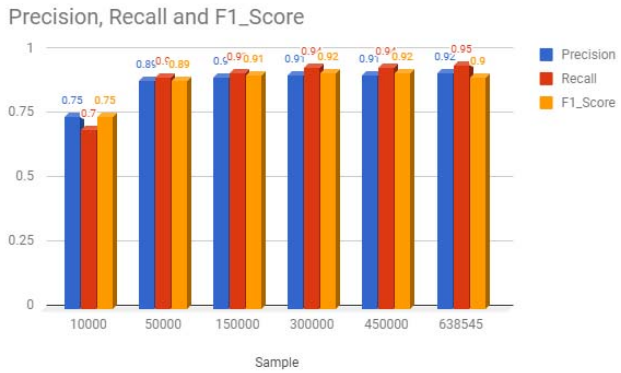

Fig. 8. Performance comparison of Precision,Recall and F!_Score for Perpetrator Gender

The ratio estimated through calculation of recall in Fig 8. is found to outscore those of precision and F1 score. However, in the case of a set of 10,000 samples, The values of precision and F1 score are observed to be greater than the recall score. This can be inferred as an indication of a larger number of false negatives present in the sample set as opposed to the number of false positives predicted by the model.

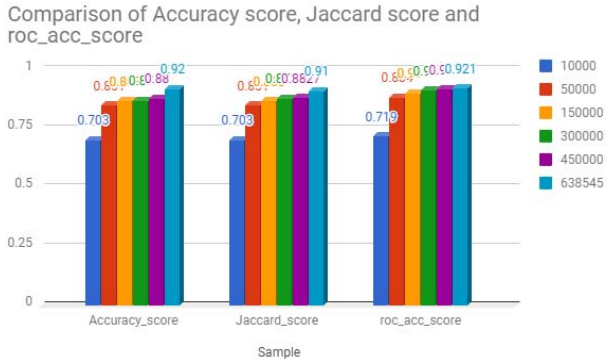

Fig. 9. Performance comparison of Accuracy score , Jaccard score and ROC_acc_score for Perpetrator Gender

The Accuracy and Jaccard similarity scores in Fig 9 found to be comparable for each of the sample sets. A common trend that can be observed from the above graph is the increase in score with respect to the increase in the size of the sample set. For example, the lowest observed ROC accuracy score was found to belong to the sample set with the smallest size (10,000 samples). Likewise, the highest score was found to be corresponding to the sample set with largest size (638454 samples).

Fig. 10. Prediction Loss Analysis of Perpetrator Gender

. From the above graph, it is infered that loss is found to decrease steadily with increase in the size of the sample set. For example, the Hamming loss observed with 10,000 samples taken into consideration is 0.22. The same metric gives a value of 0.06 with a sample set of 6,38,454 samples.

3) Performance Analysis of Perpetrator Relationship using K-Neighbors Classifier

The Performance analysis of Perpetrator Relationship is done through Precision, Recall, F1_Score, ROC, Accuracy score and Jaccard Score.

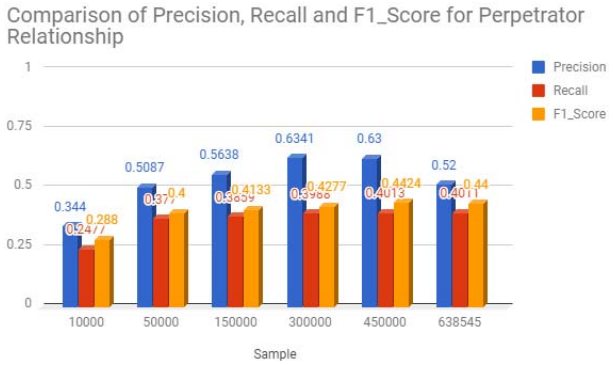

Fig. 11. Performance comparison of Precision,Recall and F!_Score for Perpetrator Relationship

Contrary to the results observed in the case of perpetrator gender, the ratio estimated through calculation of precision in Fig 11 is found to significantly outscore those of recall and F1 score. The trend found in Fig 12 is persistent for all sample sets. However, the Accuracy and Jaccard similarity scores are scores are once again found to be comparable for each of the sample sets. Similarly, in the case of ROC accuracy score, the accuracy increases with the increase in the sample set.

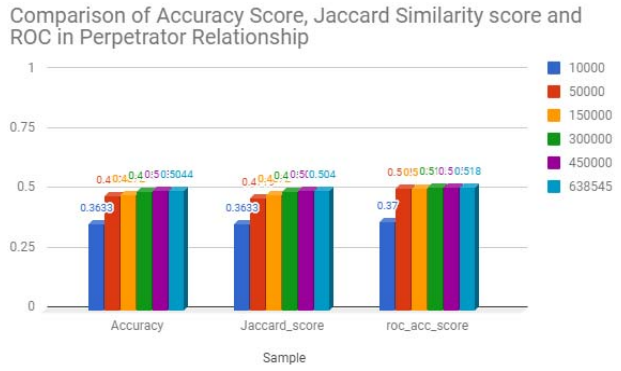

Fig. 12. Performance comparison of Accuracy score, Jaccard score and ROC_acc_score for Perpetrator Relationship

Fig. 13. Prediction Loss Analysis of Perpetrator Gender

In the above graph, , the Hamming loss observed with 10,000 samples taken into consideration is 0.03. The same metric gives a value of 0.02 with a sample set of 6,38,454 samples and the zero_one_loss decreases drastically.

4) Performance Analysis of Perpetrator Gender and Perpetrator Relationship using Neural Networks

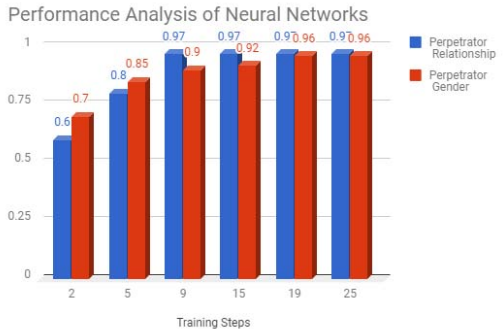

Fig. 14. Performance Analysis of Neural Network

From the above graph Fig 14, it is inferred that, the accuracy score increases with the increase in the training steps. Using Neural Networks, Gender is predicted with an accuracy of 0.96 and Relationship is predicted with an accuracy of 0.97.

V. Conclusion and Future Work

On implementing this model, it produces an accuracy of 0.60 for Perpetrator Age, 0.96 for Perpetrator Sex and 0.97 for Relationship. The Accuracy of the system can be improved by using complex neural networks like Concurrent Neural Networks and Recurrent Neural Networks.

Acknowledgment

We thank faculties from Department of Computer Science and Engineering, Loyola-ICAM College of Engineering and Technology for their valuable support and guidance.

References

[1]. Chen, Ling, and Xu Lai. "Comparison between ARIMA and ANN models used in short-term wind speed forecasting." *Power and Energy Engineering Conference (APPEEC), 2011 Asia-Pacific*. IEEE, 2011.

[2]. Agarwal, Jyoti, Renuka Nagpal, and Rajni Sehgal. "Crime analysis using K-means clustering." *International Journal of Computer Applications* 83.4 (2013).

[3]. Sathyadevan, Shiju, and Surya Gangadharan. "Crime analysis and prediction using data mining." *Networks & Soft Computing (ICNSC), 2014 First International Conference on*. IEEE, 2014

[4]. McClendon, Lawrence, and Natarajan Meghanathan. "Using machine learning algorithms to analyze crime data." *Machine Learning and Applications: An International Journal (MLAIJ)* 2.1 (2015).

[5]. Kiani, Rasoul, Siamak Mahdavi, and Amin Keshavarzi. "Analysis and prediction of crimes by clustering and classification." *Analysis* 4.8 (2015).

[6]. Heartfield, Ryan, George Loukas, and Diane Gan. "You are probably not the weakest link: Towards practical prediction of susceptibility to semantic social engineering attacks." *IEEE Access* 4 (2016): 6910-6928.

[7]. Sivaranjani, S., S. Sivakumari, and M. Aasha. "Crime prediction and forecasting in TamilNadu using clustering approaches." *Emerging Technological Trends (ICETT), International Conference on*. IEEE, 2016.

[8]. Kansara, Chirag, et al. "Crime mitigation at Twitter using Big Data analytics and risk modelling." *Recent Advances and Innovations in Engineering (ICRAIE), 2016 International Conference on*. IEEE, 2016.

[9]. Tsunoda, Masateru, Sousuke Amasaki, and Akito Monden. "Handling categorical variables in effort estimation." *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*. ACM, 2012.

[10]. Su, Ya, et al. "Multivariate multilinear regression." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.6 (2012): 1560-1573.

[11]. Viswanath, P., and T. Hitendra Sarma. "An improvement to k-nearest neighbor classifier." *Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE*. IEEE, 2011.

[12]. Palocsay, Susan W., Ping Wang, and Robert G. Brookshire. "Predicting criminal recidivism using neural networks." *Socio-Economic Planning Sciences* 34.4 (2000): 271-284