Crime in India against Women

Pratham Saraf

Manipal University Jaipur

**Abstract—This paper helps to predict future crime rates and classify states as safe or unsafe. The datasets used in the paper record the Indian crime history from 2001-2015.Several prediction and classification techniques have been used to analyze the following data. Linear Regression, SVM, Naive Bayes and 4 other algorithms have been implemented.**

I. Introduction

Crime is a curse that haunts mankind forever, it affects the economic and social growth of a nation and when the nation comprises of a massive population of 1.4billion people the crisis becomes even worse. With diverse culture and varied communities, the crime in India varies from state to state and city to city. The major factors that contribute to the rise in criminal activities are education, poverty, and employment. The crime rates have seen a significant drop in the country but still the condition is alarming. Crime against women in still very prevalent in most of the parts of the country.

According to India's National Crime Records Bureau, reported incidents of violence against women increased by 6.4% in 2012, with reports of crimes against women occurring every three minutes. According to the National Crime Records Bureau, there were more than 228,650 incidents of violence against women in 2011 and more than 300,000 in 2015, a 44% increase. 7.5% of women living in India live in West Bengal, accounting for 12.7% of all reported crimes against women. Andhra Pradesh is home to 7.3% of Indian women and accounts for 11.5% of all reported female cases.

Machine learning in this field would help to analyse the past rate of crimes and predict the future rates of crime. The data used was collected from Kaggle and then was later processed. The data was then grouped by states and year for analysis and several algorithms were further applied.

II.Data Analysis

1. Data Source

The data that was downloaded from the Kaggle was disarranged and needed to be grouped. The data was then grouped in two ways, one according to the year format and the other was according to state names. The dataset provides information about the various types of crimes against women. One more dataset used was of the people getting convicted which was grouped according to the state names. These two crime datasets were used for prediction and classification.

III. Machine Learning

Machine learning is the part of artificial intelligence which deals with statistical methods and gives computers the ability to learn from past experiences [28]. Machine learning can be divided into categories such as supervised, unsupervised, and reinforcement learning. This study uses supervised learning and unsupervised learning both.

Supervised learning can be categorized into classification and regression. Classification is predicting a discrete class label, while regression is the task of predicting a continuous quantity. This work attempts to predict the types of crime in a particular location. Therefore, the objective of this study is the classification of crime. There are many algorithms that can be used for the classification such as K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Naïve Bayesian, Decision Tree, and Ensemble Methods. Every algorithm has its own advantage and disadvantage in terms of complexity, accuracy, and training time and can provide different results from a single dataset. Unsupervised learning on the other hand includes methods such as PCA and K-Means. These methods are generally used for clustering purpose and the predictions are based on a similar approach.

The algorithms that have been implemented on the dataset include Linear Regression, KNN, K- Means, Naive Bayes, SVM also PCA and CovMatrix were performed.

Approach 1:

This approach involved the application of linear regression. Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e., it finds the linear relationship between the dependent and independent variable. A Linear Regression model’s main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.

The data was grouped by years and the total crimes against women was used as the prediction set. The other data was used to make the prediction. The model made had an accuracy of 99% and can be used to predict the crime rates for future years. The library used for the Linear Regression model was the Sklearn.linear\_model and the model selection library were used to split the data.

Chart, scatter chart

Description automatically generated

Approach 2:

This approach involved the application of gaussian Naïve Bayes. The Naive Bayes classifier greatly simplify learning by assuming that features are independent given class. Although independence is generally a poor assumption, in practice naive Bayes often competes well with more sophisticated classifiers. This method is generally used for classification predictions. It has great accuracy while predicting Boolean results but not much for pure number predictions.

Still, we attempted to use Naïve Bayes as one of the methods for the prediction and as the accuracy calculation is strictly based on the correct or incorrect result another method had to be employed to find the accuracy. Which was found to be pretty accurate and was 95.7%. The method used was somewhat related to variance and mean for checking the closeness of the values to test values.

Approach3:

This approach involved the application of K- Means. The division-based KMeans algorithm is a variant of the clustering algorithm proposed by J. B. McQueen. This unsupervised algorithm is commonly used for data mining and pattern recognition. This algorithm is based on cluster figure of merit, squared error, and the desire to minimize error criteria. To find the optimization result, the algorithm tries to find K partitions that meet certain criteria. First, we select a few points that represent the initial focus of the cluster (usually we choose the first K-sample income points to represent the initial focus of the cluster). Second, according to the minimum distance criterion, we obtain an initial classification to collect the remaining sample points as foci, and if the classification turns out to be absurd, we correct (recalculate each foci of the cluster) and repeat until we obtain a reasonable classification.

Two important attributes which were rape and assault on women with intent to outrage her modesty were used to form clusters. There data was compared simultaneously and later plotted on the graphs to perform the K-means distribution. Four clusters were mainly formed which included one of the clusters to be very small and could be excluded as a outlier.

Chart, scatter chart

Description automatically generated

Approach 4:

This approach involved the application of KNN. kNN classifier is to classify unlabelled observations by assigning them to the class of the most similar labelled examples. Characteristics of observations are collected for both training and test dataset. The minowski distance was used to calculate the K- nearest Neighbours. Since the kNN is also widely used for classification purposes and the method to calculate accuracy is quite similar to the one used by naïve bayes. We implemented another method to get more accurate results. For the accuracy we calculated the variance of the test and prediction values and then the mean to find an average accuracy for the total number of cases.

Approach 5:

Principal component analysis was also performed on the dataset after classifying the states as unsafe and safe. The classification was made on the basis of people that were convicted and total people that were arrested. Another column was added to the dataset which listed the states as safe or unsafe. The PCA helped to find the main components which influenced the data the most.

Approach 6:

This approach involved the application of HAC. The Classical Agglomerative Clustering (HAC) algorithm is widely used in modern data science and creates hierarchical relationships between data samples by partitioning a data set into clusters. HAC algorithms are used in many applications such as biology, natural language processing, and recommendation systems. Therefore, it is necessary to verify that these algorithms are valid. Even if the data set contains biases for specific protection groups, the generated cluster output should not distinguish between samples from these groups. The two factors that were used were the total number of cases and the number of rape cases to plot the dendrogram for further analysis. It was seen that there were 4 major clusters being formed.

Chart, histogram

Description automatically generated

Approach 7:

Some algorithms that have been mentioned above were used again by grouping the datasets differently. The naïve bayes algorithm was also implemented on the Safety Status dataset. Which resulted in a model that provided results with an accuracy of 81.8%. Similarly linear regression was also used to predict the crime rates solely dependent on the states but due to data being insufficient overfitting was experienced and the model had an accuracy of 100% which basically rendered it to be useless.

V. Conclusion

In this research, Indian crime data for the last 15 years was used in two different dataset approaches. Machine learning predictive models Naïve Bayes, KNN, K-Means, SVM, PCA, CovMatrix and Linear Regression were used to obtain crime-prediction accuracy between 80% to 95%. The accuracy, complexity, and training time of algorithms were slightly different for different approaches and algorithms. The prediction accuracy can be improved by taking a larger data set aur by breaking down the existing categories.

Acknowledgment

This project was based on crime in India against women which was assigned in the 4th semester under the machine learning subject. I would like to thank my professor Mr. Venkatesh Gauri Shankar for the support provided throughout project journey.

References

[1] [https://www.researchgate.net/profile/Pooya- Taheri/publication/330475412\_Crime\_Analysis\_Through\_Machine\_Learning/links/5c467bba92851c22a386ffbd/Crime-Analysis-Through-Machine-Learning.pdf](https://www.researchgate.net/profile/Pooya-%20%20%20Taheri/publication/330475412_Crime_Analysis_Through_Machine_Learning/links/5c467bba92851c22a386ffbd/Crime-Analysis-Through-Machine-Learning.pdf)

[2] [www.Kaggle.com](http://www.Kaggle.com)

[3] <https://jovian.ai/sathi-satb/crime-against-women>

[4] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4916348/>

[5] <https://www.sciencedirect.com/science/article/pii/S1875389212006220>

[6] <http://www.womenlawsindia.com/legal-awareness/crimes-against-women/>

[7] <https://scikit-learn.org/stable/modules/naive_bayes.html>

[8] <https://ieeexplore.ieee.org/abstract/document/8203676>

[9] <http://ijmore.co.in/index.php/ijmore/article/view/5>

[10[] http://ijmore.co.in/index.php/ijmore/article/view/5](%5d%20http:/ijmore.co.in/index.php/ijmore/article/view/5)