Estimation of crime prediction using K-Nearest Neighboring algorithm of machine learning.

Akash Kumar[α]; Aniket Verma[β]; Gandhali Shinde[γ]
Yash Sukhdeve[ζ]; Nidhi Lal[σ]
Department of Computer Science and Engineering
IIIT Nagpur,
Nagpur-4110001, India
[α]akkshroy@gmail.com
[β]aniketverma98@gmail.com
[γ]gandhalishinde27@gmail.com
[ζ]sukhdeveyash@gmail.com
[σ]nidhi.lal@cse.iiitn.ac.in

***Abstract*— For a developing country like India, it is not new that people hear of crimes happening quite often. With the rapid urbanization of cities, we have to constantly be aware of our surroundings. In order to avoid the unfortunate, we will try to analyze crime rates by the KNN prediction method. It will predict, tentatively, the type of crime, when, where and at what time it may take place.**

This data will give the behaviors in crime over an area which might be helpful for criminal investigations. It will also provide us with the most committed crime in a particular region. In this paper, we will use the k-nearest neighbor algorithm of machine learning.

***Keywords*: Data Analysis ; Crime Prediction ; Machine Learning ; K-Nearest Neighbors.**

I. INTRODUCTION

Criminal activity is gradually rising in India and has a significant and negative social impact[3]. The recent spurt in the nation has put everyone wondering as to what will happen in the future. Cases of murder, abduction, rape, and fatal accidents have skyrocketed. The need of the hour is to make people of the nation realize the issue. Machine learning advancements and deep learning algorithms can discover hidden patterns in unstructured data sets and reveal new information. Crime prediction and criminal identification are the major problems to the police department because there is a tremendous amount of data related to crime that exists. There is a need for technology through which the case-solving could be faster[3]. The idea behind this project is that crimes can be easily predicted once we are able to sort through a huge amount of data to find patterns that are useful to configuring what is required[1]. The recent developments in machine learning makes this task possible. We will give date, time, location (longitude, latitude) as input and the output will be generated which will give us information about which crime is likely to happen in that area. It basically gives us the hotspots of crime[5]. The data is taken considering the time and type of crime that happened in the past. KNN algorithm then uses its approach which assumes that similar things exist in close proximity and classifies new cases based on similarity measures.

Classes of crimes are:

- Act 379 - Robbery
- Act 13 - Gambling
- Act 279 - Accident
- Act 323 - Violence
- Act 302 - Murder
- Act 363 - Kidnapping

This prediction, if put to good use, is of great help in investigating cases that have happened. It can be used to suppress the crimes by installing some measures if we know what type of crime is going to happen beforehand. This will indirectly help reduce the rates of crimes and can help to improve security in such required areas[2].

II. RELATED WORK:

It is observed that many machine learning models are implemented on datasets of different cities having unique

features, so predictions are different in all cases. Classification models have been implemented on various other applications like prediction of weather, in banking, finances and also in security [3].

In [4] identification of criminals by using classification techniques and crime prediction was done using data set of six cities of Tamil Nadu by using KNN classification, K-Means clustering, Agglomerative hierarchical clustering, and DBSCAN clustering algorithms. In [5], they used a model whose purpose was to use a database in which the data points were separated into several classes to predict the classification of a new sample point. Using features Day, Date, Year of the crime using KNN it is found to be 40% accuracy. Their model used techniques like Logistic Regression, Decision Trees, Bayesian Methods and Support Vector Machine[9]. Python was used to explore training data, make regression analysis and predict categories for test data, in order to get the best correlation between the features (Date, Pd-District, Address, Day of the Week, Description, Resolution, X and Y) and the target value (Category of Crime). All nominal values were converted into binary values by converting the values of the attributes into separate new attributes and give them values of either a 0 or 1. Several trials of different Regression methods were used on the training data by splitting it into two sets; training and validation, both validation and cross-validation were conducted, the method with the least Log loss was applied to predict the results for the test data.

III. PROPOSED WORK:

Processing data:

Initially, we need to preprocess data by removing all null values and columns that are unnecessary[2]. The dataset that we have used is a modification of the original dataset that was obtained by scraping the police website of a city Indore in Madhya Pradesh. It was processed multiple times and they dropped features such as police station, station number, Complainant name & address, accused name & address.

There were minor modifications that we made in their final dataset. We calculated the importance of features by Extra Trees Classifier function which helped us in neglecting the unnecessary attributes (refer Table 1).

Extra Trees Classifier is a type of ensemble learning technique which takes a specified amount of data (value of n-estimators) and calculates the importance of each feature separately.

Figure 1 shows the importance of each feature in through a bar graph.

Features	Importance
Hour	0.2961921
Latitude	0.3061108
Longitude	0.2677053
Year	0.00000
Month	0.00012
Week of the year	0.00441

Table 1: Importance of Features

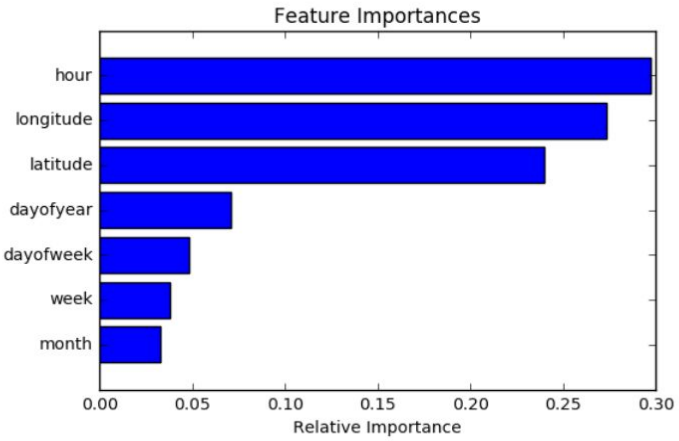

Figure 1: Importance of Features

	hour	dayofyear	act379	act13	act279	act323	act363	act302	latitude	longitude
0	21.0	59.0	1	0	0	0	0	0	22.737260	75.875987
1	21.0	59.0	1	0	0	0	0	0	22.720992	75.876083
2	10.0	59.0	0	0	1	0	0	0	22.736676	75.883168
3	10.0	59.0	0	0	1	0	0	0	22.746527	75.887139
4	10.0	59.0	0	0	1	0	0	0	22.769531	75.888772

Figure 2: Final Dataset

We dropped those attributes that had the least importance (year, month, week of the year, etc.). The final dataset (refer Figure 2) now has four attributes with hour, day of the year, longitude and latitude of the city. After the final dataset was made, we created another sub-set (refer Figure 3) by using SQL (refer Figure 4). We made an sns heatmap (refer Figure 4) to get a rough idea about how the crimes are varying with respect to days of a month. A heat map is a data analysis software that uses color the way a bar graph uses height and width as a data visualization tool. Our observations were that act 323 a.k.a violence had occurred most towards the month-end whereas act 279 a.k.a accidents happened alternatively.

	day	act	frequency
0	1	act379	121
1	1	act13	22
2	1	act279	88
3	1	act323	66
4	1	act363	33
5	1	act302	0
6	3	act379	66
7	3	act13	0
8	3	act279	121
9	3	act323	66

Figure 3: Sub Dataset for Heatmap

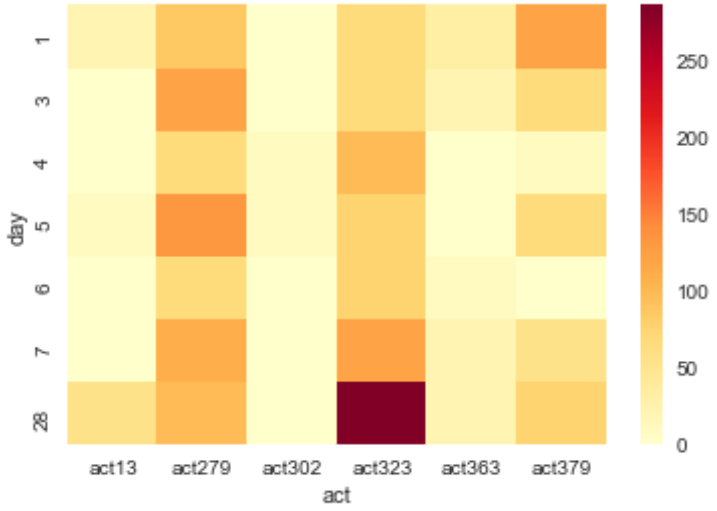

Figure 4: SNS Heatmap

The KNN Algorithm:

The next step was to decide which algorithm to use. K Nearest Neighbor Classifier is a supervised machine learning algorithm useful for classification problems. It works by finding the distances between a query and all the examples in the data, selecting the specified examples that are closest to the query, then votes for the most frequent label. It is non-parametric which means that it does not make any assumptions on the underlying data distribution. In other words, the model structure is determined by the data. It's pretty useful because in reality, most of the data does not obey the typical theoretical assumptions made[4]. Hence we decided to use K-Nearest-Neighbor Algorithm.

Applying Algorithm:

We split the train and test data in the following manner.The following pie chart (refer Figure 5) indicates:

- Training dataset consists of 80% data.
- Testing dataset consists of 20% data.

Our test set serves as a proxy for new data. We should make sure that it is representative of the data set as a whole. To predict the value of k, we plotted a graph by using the Elbow method.

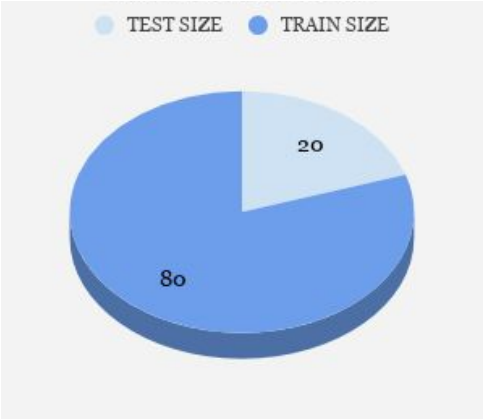

Figure 5: Split sizes

The Elbow Method is one of the most popular methods to determine this optimal value of k. The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. After analyzing the graph, the range in which the error rate was minimum came out to be 1-15. Furthermore, we checked all its values in range 1-15, but from the values of k ranging 1-13, the accuracy remained constant. Hence we selected k=3 that belonged from the range 1-13. To calculate MAE and RMSE values, we required test, train and predicted values of both x and y.

RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It’s the square root of the average of squared differences between prediction and actual observation. Both MAE and RMSE express average model prediction error in units of the variable of interest. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. RMSE avoids the use of taking the absolute value, which is undesirable in many mathematical calculations. After calculating both the values, we plotted two graphs.

$$MAE = \frac{1}{n}\sum_{j=1}^{n} |y_i - y_i|$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

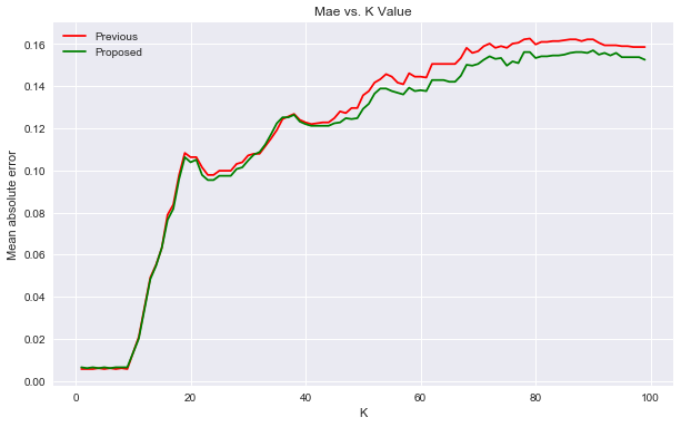

Figure 6: MAE vs. K

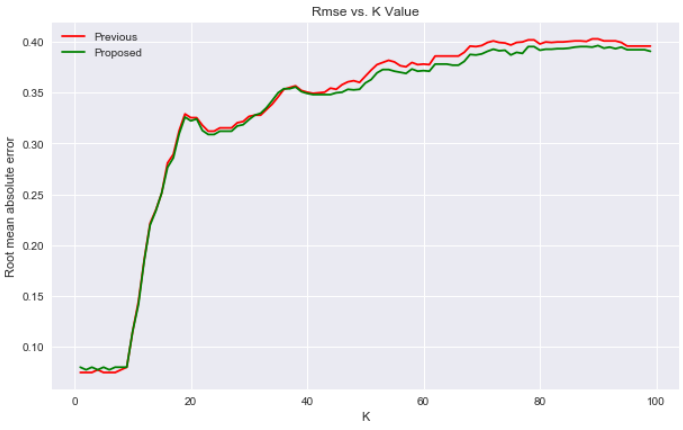

Figure 7: RMSE vs. K

The graph (refer Figure 6) indicates the Mean Absolute Error (Y-axis) and values of k (X-axis). MAE measures the average magnitude of the errors in a set of predictions, without considering their direction.

The second graph (refer Figure 7) shows Root Mean Square Error (X-axis) and values of k (Y-axis). The Root Mean Square Error is a frequently used measure of the differences between the values predicted by a model and the values observed.

The red curve indicates Previous work and the green curve indicates proposed work. Clearly, the mean absolute error and root mean square error is reduced when compared with previous work. The red curve indicates Previous work and the green curve indicates proposed work. The root mean square error is reduced when compared with previous work. Finally, after calculating all error and mean values, the accuracy score of the program was calculated. Mean Absolute Error (MAE) and Root mean squared error (RMSE) are two of the most common metrics used to measure accuracy for continuous variables.

IV. RESULTS AND DISCUSSIONS:

Solving a crime problem is a complex task that requires human experience and intelligence and also methods that can help them with crime detection problem[6]. By using historical data and observing where recent crimes took place we can predict where future crimes will likely happen.

Figure 8: Error Rate vs. K Value

	Previous	Proposed
Mean Absolute Error	0.1598	0.0064
Root Mean Square Error	0.3997	0.0802
KNN Score	0.9323	0.9951

Table 2

We chose k=3 to acquire the highest accuracy as possible.The above graph (refer Figure 8) indicates the RMSE value plotted against K value. The yellow curve is for proposed work and blue is for previous work. An increase in k-value results in increased root mean square error. Hence the value of k was picked from range 1-15 because that is the only range with minimum error. The previous work had

included extra factors which did not seem necessary in our case.

Comparison between the results from previous work and proposed work is indicated in Table 2.

As it is clear from the graph, now we can confidently say that the error was reduced thus increasing the accuracy of the program[3]. Crime patterns cannot be static since patterns change over time. By training means we are teaching the system based on some particular inputs. So the system automatically learns the changing patterns in crime after analyzing them. Also, we cannot ignore the fact that crime factors change with time[3].

V. CONCLUSION:

This research work presents a method to predict and forecast crimes within a city. It focuses on having a crime prediction tool that can be helpful to law enforcement[6]. Hence we tried to increase the prediction accuracy as much as possible. As compared to the previous work, we were successful in achieving the highest accuracy in prediction. The values of RMSE and MAE were reduced significantly[8]. Along the way, we got to know the patterns of criminal activities in various areas which will be helpful for criminal investigation. This pattern has much greater importance than we realize. The KNN system assists law enforcement agencies for improved and accurate crime analysis. By sifting through the crime data we have to identify new factors that lead to crime[3]. Since we are considering only some limited factors full accuracy cannot be achieved. For getting better results in prediction we have to find more crime attributes of places instead of fixing certain attributes[12]. Till now we trained our system using certain attributes but we can include more factors to improve accuracy. In the future, this work can be extended to have improved classification algorithms to identify criminals more efficiently[7]. We believe that the crime rates that are increasing non-stop may go down in the future due to such prediction techniques.

VI. REFERENCES:

[1] Kim, Suhong, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri. "Crime Analysis Through Machine Learning." In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 415-420. IEEE, 2018.

[2] Shah, Riya Rahul. "Crime Prediction Using Machine Learning." (2003).

[3] Lin, Ying-Lung, Tenge-Yang Chen, and Liang-Chih Yu. "Using machine learning to assist crime prevention." In 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 1029-1030. IEEE, 2017.

[4] M. V. Barnadas, Machine learning applied to crime prediction, Thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, Sep. 2016.

[5] Crime Prediction Using Machine Learning Sacramento Stateathena.ecs.csus.edu › ~shahr › progress_report by RR Shah - 2003.

[6] Shamsuddin, Nurul Hazwani Mohd, Nor Azizah Ali, and Razana Alwee. "An overview on crime prediction methods." In 2017 6th ICT International Student Project Conference (ICT-ISPC), pp. 1-5. IEEE, 2017.

[7] Sivaranjani, S., S. Sivakumari, and M. Aasha. "Crime prediction and forecasting in Tamilnadu using clustering approaches." In 2016 International Conference on Emerging Technological Trends (ICETT), pp. 1-6. IEEE, 2016.

[8] Ozgul, Fatih, Zeki Erdem, and Chris Bowerman. "Prediction of unsolved terrorist attacks using group detection algorithms." In Pacific-Asia Workshop on Intelligence and Security Informatics, pp. 25-30. Springer, Berlin, Heidelberg, 2009.

[9] Williams, Matthew L., Pete Burnap, and Luke Sloan. "Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns." The British Journal of Criminology 57, no. 2 (2017): 320-340.

[10] Ivan, Niyonzima, Emmanuel Ahishakiye, Elisha Opiyo Omulo, and Danison Taremwa. "Crime Prediction Using Decision Tree (J48) Classification Algorithm." (2017).

[11] Iqbal, Rizwan, Masrah Azrifah Azmi Murad, Aida Mustapha, Payam Hassany Shariat Panahy, and Nasim Khanahmadliravi. "An experimental study of classification algorithms for crime prediction." Indian Journal of Science and Technology 6, no. 3 (2013): 4219-4225.

[12] Agarwal, Shubham, Lavish Yadav, and Manish K. Thakur. "Crime Prediction Based on Statistical Models."

In 2018 Eleventh International Conference on Contemporary Computing (IC3), pp. 1-3. IEEE, 2018.

[13] Babakura, Abba, Md Nasir Sulaiman, and Mahmud A. Yusuf. "Improved method of classification algorithms for crime prediction." In 2014 International Symposium on Biometrics and Security Technologies (ISBAST), pp. 250-255. IEEE, 2014.

[14] Jia, Xueming. "Crime Prediction Using Data Mining and Machine Learning." In *The 8th International Conference on Computer Engineering and Networks (CENet2018)*, p. 360. Springer.

[16] Wu, Shaobing, Changmei Wang, Haoshun Cao, and Xueming Jia. "Crime Prediction Using Data Mining and Machine Learning." In *International Conference on Computer Engineering and Networks*, pp. 360-375. Springer, Cham, 2018.

[17] Yadav, Sunil, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma, and Nikhilesh Yadav. "Crime pattern detection, analysis & prediction." In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, vol. 1, pp. 225-230. IEEE, 2017.

[18] Nakaya, Tomoki, and Keiji Yano. "Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics." *Transactions in GIS* 14, no. 3 (2010): 223-239.