Crime Prediction Using K-Nearest Neighboring Algorithm

Akash Kumar
Department of Computer Science and Engineering
IIIT Nagpur,
Nagpur-4110001, India
akkshroy@gmail.com

Aniket Verma
Department of Computer Science and Engineering
IIIT Nagpur,
Nagpur-4110001, India
aniketverma98@gmail.com

Gandhali Shinde
Department of Computer Science and Engineering
IIIT Nagpur,
Nagpur-4110001, India
gandhalishinde27@gmail.com

Yash Sukhdeve
Department of Computer Science and Engineering
IIIT Nagpur,
Nagpur-4110001, India
sukhdeveyash@gmail.com

Nidhi Lal
Department of Computer Science and Engineering
IIIT Nagpur,
Nagpur-4110001, India
nidhi.lal@cse.iiitn.ac.in

2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) 978-1-7281-4142-8/20/$31.00 ©2020 IEEE 10.1109/ic-ETITE47903.2020.155

***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 observe 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.

Key Data Analysis; Crime Prediction; Machine Learning; K-Nearest Neighbors component, formatting, style, styling, insert (*key words****)***

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 find new patterns in various data sets and reveal new information. Crime prediction and identifying criminals are the one of the top priority 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. One 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 seen that many of machine learning models are built on datasets of different cities having different unique features, so assumption is 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 main aim was to use a dataset where the data positions were divided into different classes to get clarity of a new sample positions. Using features like Day, Date, Year of the crime using KNN - algorithm it is found to be 40% accurate. Their model used techniques like Logistic Regression, Decision Trees, Bayesian Methods and Support Vector Machine [9]. Python came into use for training data, make regression analysis and conclude the categories for test dataset, to get the best correlation between the features (Hour, Longitude, Latitude, Day of the Week, Week, Month) and the destination value (Divisions of Crime). All-important values were changed into binary values by making the values of the features values into separate new attributes and convert values into either a 0 or 1. There were many trails of different Regression methods were used on the training dataset by splitting it into two sets; training and testing, both validation and cross-testing were conducted, the method with the lowest loss was applied to get the results for the test data.

III. PROPOSED WORK

A. Processing data:

Initially, data was preprocessed by removing all null values and columns that are unnecessary [2]. The dataset that was 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 made in their final dataset. importance of features was calculated by the 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 and every feature separately.

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

TABLE 1: Importance of Features

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

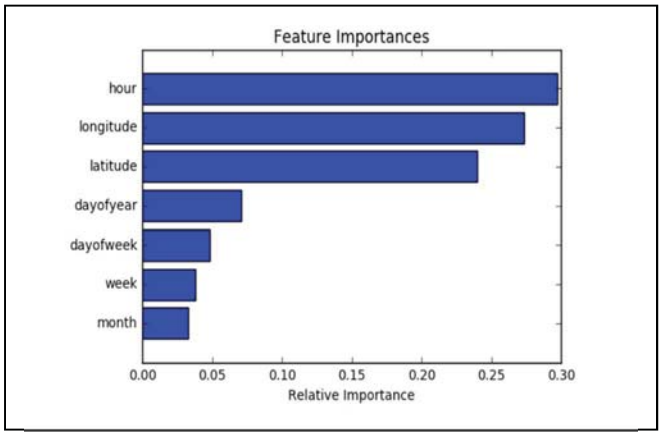

Figure 1: Importance of Features

	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

The attributes of least importance were dropped (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, another sub-set (refer Figure 3) was created by using SQL (refer Figure 4). A SNS heatmap was generated (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 like data analysis software which takes the help of colors the way similar to a bar graph which uses height and width as a data visualization tool. [8] Our observations were that act 323 i.e. violence had occurred most towards the month-end whereas act 279 i.e. 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

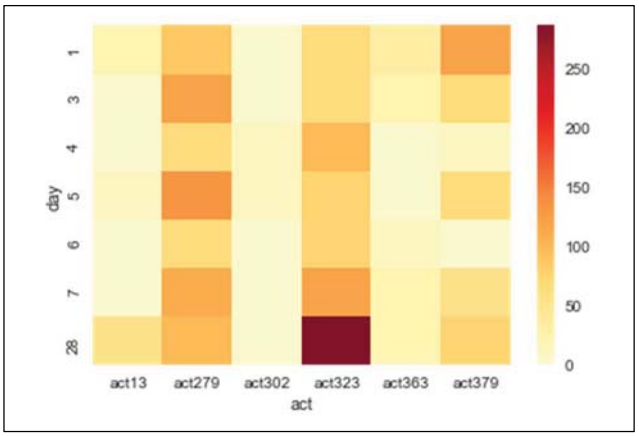

Figure 4: SNS Heatmap

B. 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, and then votes for the most frequent label. It is not parametric which implies that it does not make any supposition on the primary data distribution. To put it in simple words, the model structure is decided by the data. It's pretty useful because in reality, most of the data does not follow the typical theoretical norms made [4]. Hence, we decided to use K-Nearest-Neighbor Algorithm.

C. Applying Algorithm:

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

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

This test set serves as a proxy for new data. One should make sure that it is representative of the data set as a whole. To predict the value of k, a graph was plotted by using the Elbow method. [7]
Our test set serves as a method.

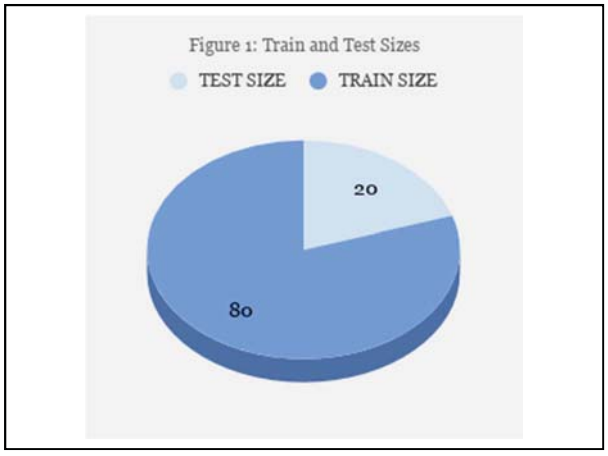

Figure 5: Split sizes

The Elbow Method is very widely used method which helps in determining the optimal value of k. The elbow method runs k-means clustering on the dataset for a variety of values for k (e.g. 1-15) and then for every value of k, it works out 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 which figures the average magnitude of the error. It is the square root of the square differences calculated between expectation and actual observation. MAE and RMSE both express average model prediction error in units of the interest variable. Provided that the errors are squared until they are averaged, large errors are assigned a fairly high weight. RMSE avoids making use of absolute value, which is unwanted in many mathematical calculations. After calculating both the values, we plotted two graphs.

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

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

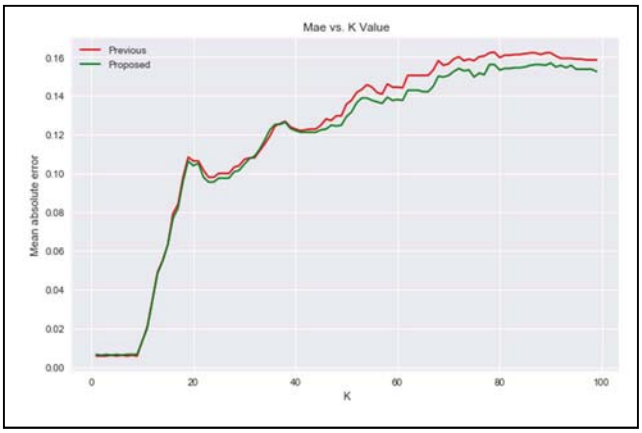

Figure 6: MAE vs. K

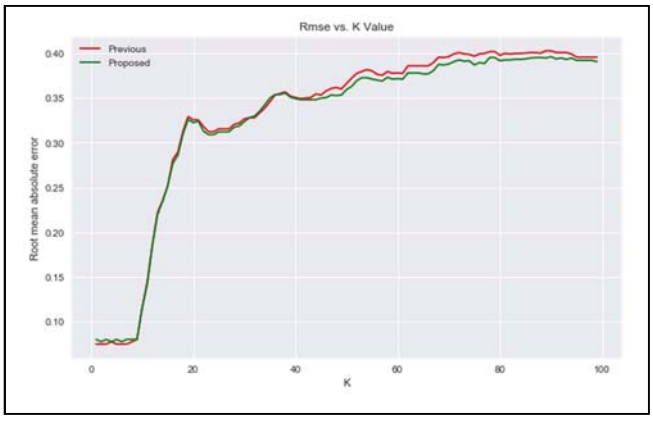

Figure 7: RMSE vs. K

The graph (refer Figure 6) indicates the Mean Absolute Error (Y-axis) and values of k (X-axis). MAE measures the

average amount of the errors in a set of predictions, without bearing in mind the direction.

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 commonly used measure of the differences between the expected values by a model and the observed values.

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. RMSE is reduced when compared with previous work.

Finally, after calculating all error and mean values, the accuracy score of the program was calculated. To calculate precision for continuous variables, Mean Absolute Error (MAE) and Root mean squared error (RMSE) are two of the most common metrics that are utilized.

IV. RESULTS AND DISCUSSIONS

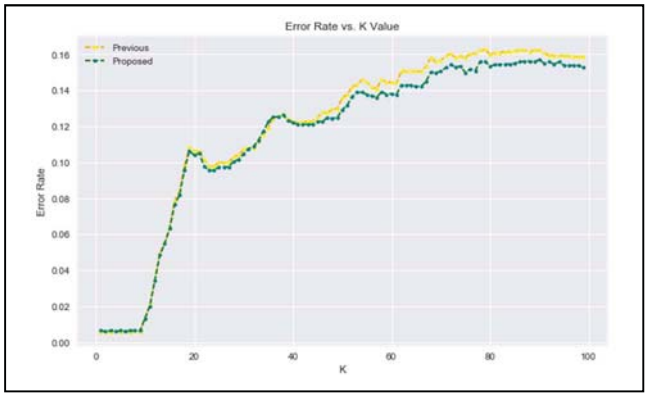

Figure 8: Error Rate vs. K Value

TABLE 2: Result

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

K = 3 helped 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 one can confidently say that the error was reduced thus increasing the accuracy of the program [3]. The patterns of crime are not same every time patterns always changes after time to time. The system was trained to learn using some particular inputs. So, the method by itself learns different changes that come in the pattern of crime after analyzing them. Also, we cannot ignore the fact that crime factors change with time [3].

V. CONCLUSION

This research work offers a way to foresee and predict crimes and frauds within a city. It focuses on having a crime prediction tool that can be helpful to law enforcement. This paper is aimed at increasing the prediction accuracy as much as possible. As compared to the previous work, this work was successful in achieving the highest accuracy in prediction. The values of RMSE and MAE were reduced significantly. Along the way, many patterns of criminal activities in various areas which will be helpful for criminal investigation were known. This pattern has much greater importance than we realize. The KNN system helps law implementing agencies for improved and exact crime analysis. By traversing through the crime dataset, we have to find out different reasons that lead to crime. Since this paper is bearing in mind only some limited factors, full accuracy cannot be accomplished. For getting more accurate results in prediction we have to find out more crime attributes of places instead of setting only certain attributes. Thus far this system was trained using certain attributes but we can take account of more factors to improve accuracy. In the future, this work can be stretched to have developed classification algorithms to detect criminals more efficiently. 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] 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.

[7] 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.

[8] 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.