iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.42898

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

A Design to Predict and Analyze Crime

Koshe Ahana Hemant[1], Mankani Neha Haresh[2], Sayyed Shafiya Majid[3]
[1, 2, 3]*Computer Engineering, Keystone School of Engineering, Pune, Maharashtra ,India*

Abstract: ***Crime is one of the dominant and alarming aspect of our society. Over the past few years, the crime rate across globe has increased exponentially. So, preventing the crime from occurring is a vital task. In the recent time, it is seen that artificial intelligence has shown its importance in almost all the field and crime prediction is one of them. However, it is necessary to maintain a proper database of the crime that has occurred. The ability to predict the crime on the basis of time, location and so on which can occur in future can help the law enforcement agencies in preventing the crime before it occurs from a strategical perspective. However, predicting the crime accurately is a challenging task because crimes are increasing at an alarming rate. Thus, the crime prediction and analysis methods are very important to detect the future crimes and reduce them. In Recent time, many researchers have conducted experiments to predict the crimes using various machine learning methods and particular inputs. For crime prediction, KNN, K-means and Random Forest and some other algorithms are used. Our system can predict regions which have high probability for crime occurrence and can visualize crime prone areas. The main purpose is to highlight the worth and effectiveness of machine learning in predicting violent crimes occurring in a particular region in such a way that it can be used by police to reduce crime rates in the society.***

.

I. INTRODUCTION

Crime is increasing considerably day by day. Crime is among the main issues which is growing continuously in intensity and complexity. Crime patterns are changing constantly because of which it is difficult to explain behaviours in crime patterns. Crime is classified into various types like kidnapping, theft murder, rape etc. The law enforcement agencies collects the crime data information with the help of information technologies (IT). But occurrence of any crime is naturally unpredictable and from previous searches it was found that various factors like poverty, employment affects the crime rate . It is neither uniform nor random. With rapid increase in crime number, analysis of crime is also required. Crime analysis basically consists of procedures and methods that aims at reducing crime risk. It is a practical approach to identify and analyse crime patterns. But, major challenge for law enforcement agencies is to analyse escalating number of crime data efficiently and accurately. So it becomes a difficult challenge for crime analysts to analyse such voluminous crime data without any computational support. A powerful system for predicting crimes is required in place of traditional crime analysis because traditional methods cannot be applied when crime data is high dimensional and complex queries are to be processed. Therefore a crime prediction and analysis tool were needed for identifying crime patterns effectively. This paper introduces some methodologies with the help of which it can be predicted that at what place and time which type of crime has a higher probability of occurrence. Ever Since the Coronavirus Pandemic hit the globe like a shockwave, multiple global data prediction systems have shown unexpected amount of fluctuations. Given the circumstances, it is essential to make accurate predictions based on data analysis.

One of the primary advantages of big data analytics software is that it can evaluate huge quantities of data much faster than humans can, plus spot trends they'd likely miss. So, from a crime-solving point of view, data analytics could help catch criminals who are trying to evade arrest.

II. METHODOLOGY

Predictive modeling was used for making predictions since it has the method which is able to build a model and has the capability to make predictions. This method consists of different algorithms of Machine Learning that can study properties from the data used for training which is used for producing predictions. It is split in two major classes one is Regression and other is classification of patterns. Regression models are based upon analysis of the relationship that are present between trends and variable in order to make predictions about the continuous variables. Whereas,the job of classification is to assign a particular class labels to a data value as output of the prediction. Division of pattern classification is in two ways i.e., Supervised and Unsupervised learning. It is already known in supervised learning that which class labels are to be used for building classification models. In unsupervised learning, these class labels are not known.

A. Working Of Project

Predictive modeling was used for making predictions since it has the method which is able to build a model and has the capability to make predictions. This method consists of different algorithms of Machine Learning that can study properties from the data used for training which is used for producing predictions. It is split in two major classes one is Regression and other is classification of patterns. Regression models are based upon analysis of the relationship that are present between trends and variable in order to make predictions about the continuous variables. Whereas, the job of classification is to assign a particular class labels to a data value as output of the prediction. Division of pattern classification is in two ways i.e., Supervised and Unsupervised learning. It is already known in supervised learning that which class labels are to be used for building classification models. In unsupervised learning, these class labels are not known. This paper deals with supervised learning[1].

1) *Data Collection:* Crime dataset from kaggle is used in CSV format.
2) *Data Preprocessing:* Data pre-processing basically involves methods to remove the infinite or null values from data which might affect the performance of the model. In this step the data set is converted into the understandable format which can be fed into machine learning models. The categorical attributes (Location, Block, Crime Type, Community Area) are converted into numeric using Label Encoder. The date attribute is splitted into new attributes like month and hour which can be used as feature for the model.
3) *Feature Selection:* Features selection is done which can be used to build the model. The attributes used for feature selection are Offense ID, District ID, Location, X coordinate , Y coordinate, Latitude , Longitude, Hour and month.
4) *Building and Training Model:* After feature selection location and month attribute are used for training. The dataset is divided into pair of x train, y train and x test, y test. The algorithms model is imported from sklearn. Building model is done using model. Fit (x train, y train).
5) *Prediction:* After the model is build using the above process, prediction is done using model.predict(x test). The accuracy is calculated using accuracy, score imported from metrics.
6) *Visualization:* Using matplotlib library from sklearn. Analysis of the crime dataset is done by plotting various graphs.

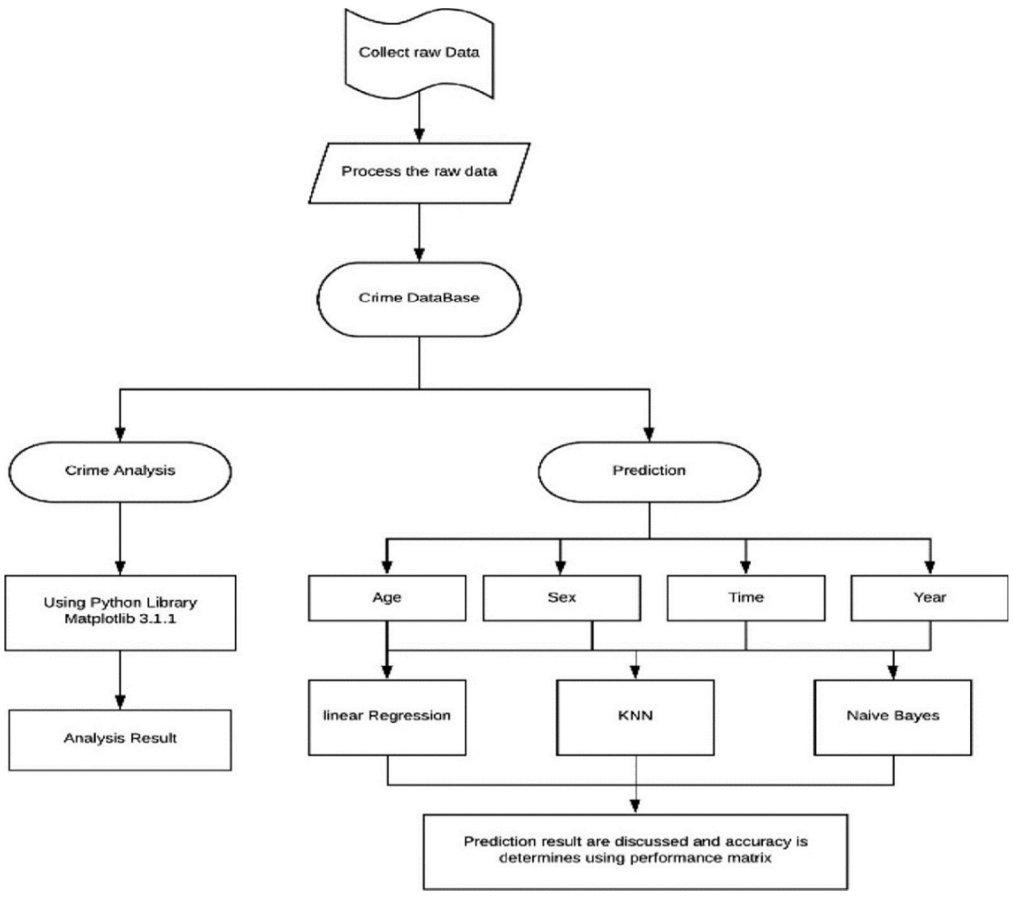


Fig 1: Methodology.

III. DATASET

The dataset that we have used is of 5 Denver City. The different columns or as we refer to them as features are as followed

```
In [2]: df = pd.read_csv('crime.csv')
        print(df.shape)
        df.head()
```

```
(462101, 19)
```

Out[2]:

	incident_id	offense_id	OFFENSE_CODE	OFFENSE_CODE_EXTENSION	OFFENSE_TYPE_ID	OFFENSE_CATEGORY_ID	FIRST_OCCURRENCE_DATE
0	20226000193	20226000193299900	2999	0	criminal-mischief-other	public-disorder	1/4/2022 11:30:00 AM
1	20223319	20223319299900	2999	0	criminal-mischief-other	public-disorder	1/3/2022 6:45:00 AM
2	20223093	20223093299900	2999	0	criminal-mischief-other	public-disorder	1/3/2022 1:00:00 AM
3	20224000	20224000299900	2999	0	criminal-mischief-other	public-disorder	1/3/2022 7:47:00 PM
4	20223956	20223956299900	2999	0	criminal-mischief-other	public-disorder	1/3/2022 5:06:00 PM

Here the values are as followed :-

1) Incident ID - Incident ID is a unique identifier for a specific crime event that happened on a specific date and time.
2) Offense ID - It indicates a unique identifier for a crime with respect to incident id and offense code(offense id= incident id + offense code)
3) Offense Code - It is a code that uniquely identifies each crime.
4) Offense Type ID - It is a superset of offense category id and indicates type of crime.
5) Offense Category ID - It is the subset of offense type id and it specifies exact details about the type of crime.
6) First Occurence Date - Date on which the particular crime was observed for the first time in the specific address.
7) Last Occurence Date - Date on which the particular crime was observed for the last time in the specific address
8) Reported Date - Date on which particular crime that took place was reported.
9) Incident Address - Address at which the crime took place.
10) GEO_LON - Longitudinal co-ordinates of the location.
11) GEO_LAT - Latitudinal co-ordinates of the location.
12) District ID - Numerical value to identify a district in which the crime took place.
13) Precinct ID - It is a numerical value or code for a police station of the area where the crime was reported.
14) Neighborhood ID - Province in the city where the crime took place.

IV. PROPOSED SYSTEM

A. *System Components*

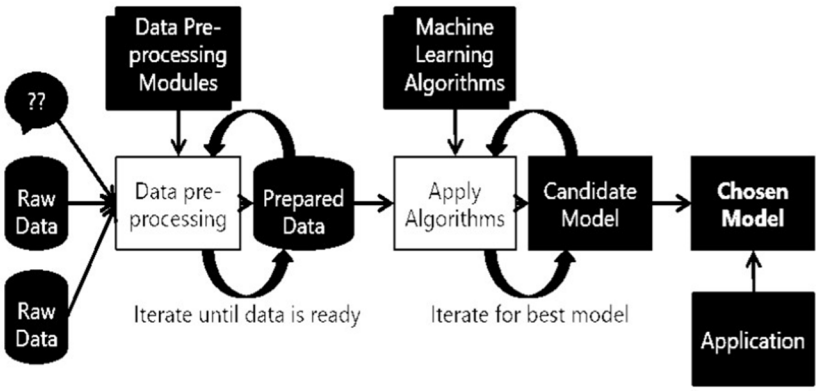


Fig 2: Flow of project.

B. System Chart

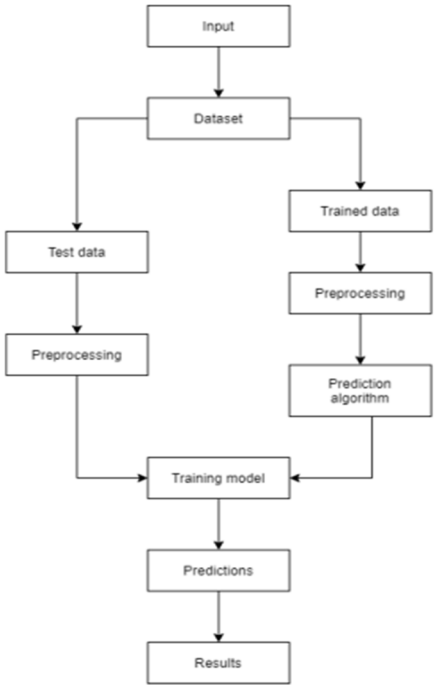


Fig 3: System Chart

C. Collaboration Diagram

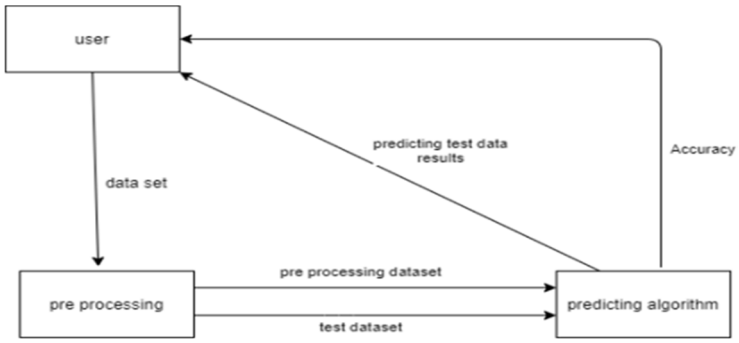


Fig 4: Collaboration Diagram

V. EDA AND DATA PRE-PROCESSING

A. Data Cleaning

The dataset is not in the right format to use directly with the model hence data preprocessing is required . In our model we did a few steps as followed :

1) Dropping the null rows

```
In [5]: df.drop(['offense_id', 'GEO_X', 'GEO_Y', 'LAST_OCCURRENCE_DATE'], axis=1, inplace=True)
```

2) Changing the data types of the column / features .

```
# feature engineering
df['FIRST_OCCURRENCE_DATE'] = pd.to_datetime(df['FIRST_OCCURRENCE_DATE'])
df['YEAR'] = df['FIRST_OCCURRENCE_DATE'].dt.year
df['MONTH'] = df['FIRST_OCCURRENCE_DATE'].dt.month
df['DAY'] = df['FIRST_OCCURRENCE_DATE'].dt.day
df['HOUR'] = df['FIRST_OCCURRENCE_DATE'].dt.hour
```

3) Sorting the data by offense category id

```
df['OFFENSE_CATEGORY_ID'].value_counts()[:15].sort_values(ascending=True).plot(kind='bar', figsize=(15,8),
                                                                                title='Bar Plot')
ax.set_xlabel("OFFENSE_CATEGORY_ID")
ax.set_ylabel("Frequency")
```

B. EDA

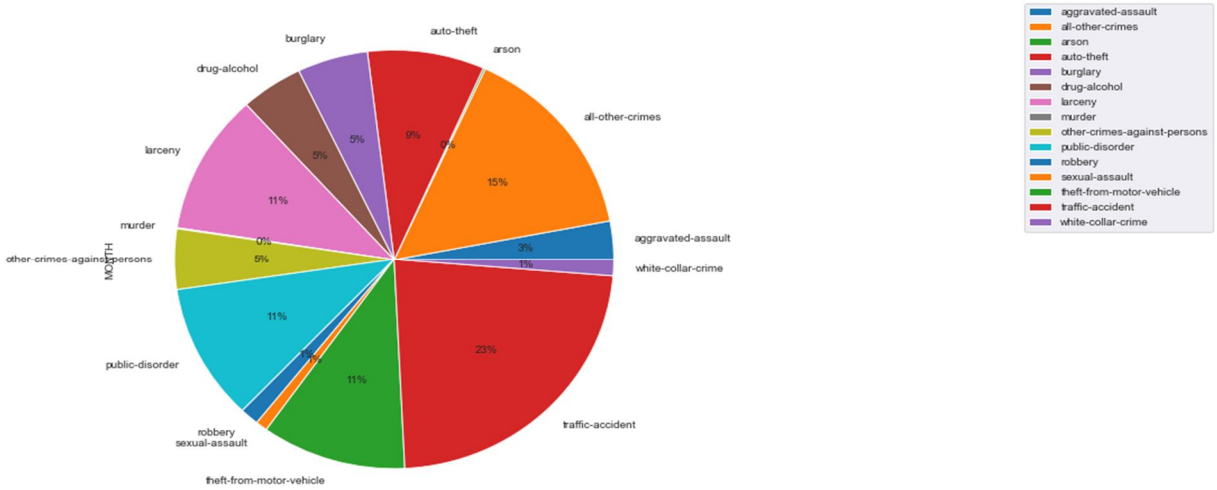


Figure 5: Percentage share of all crimes

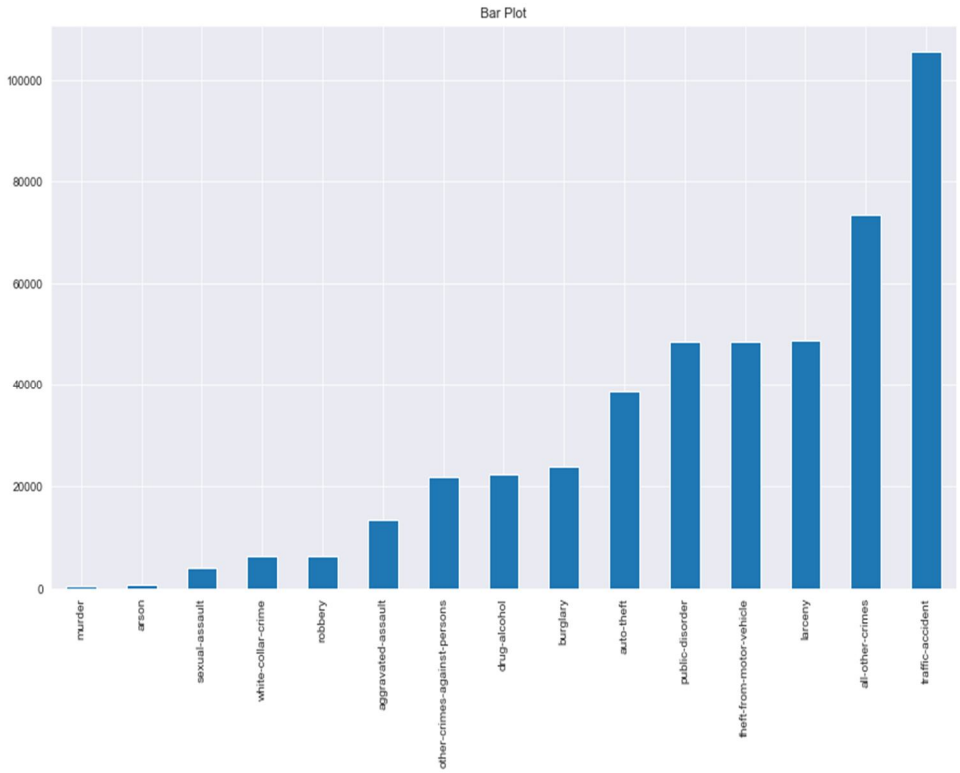


Figure 6: - Offense Category Id Vs Frequency

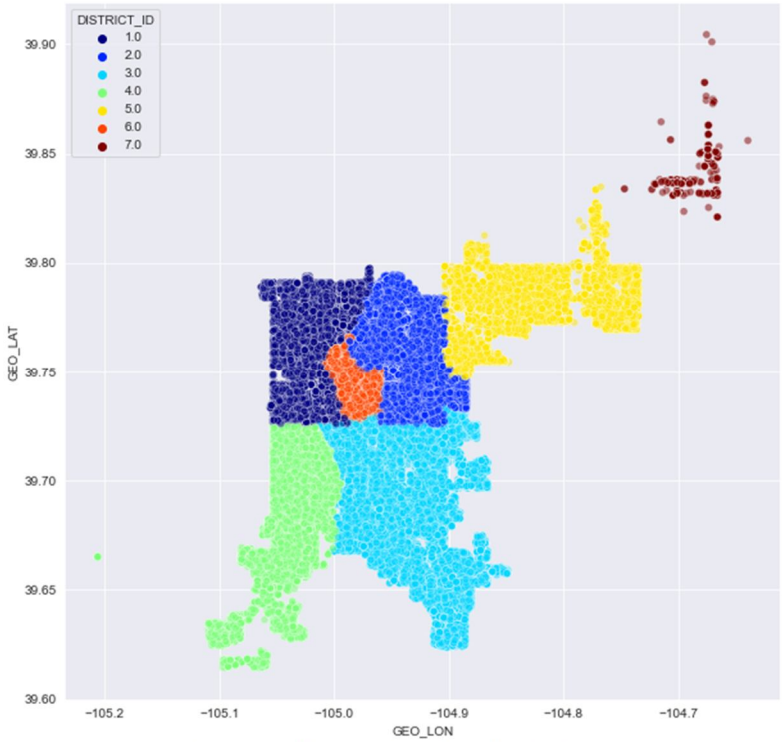


Figure 7: Color coding OD district id

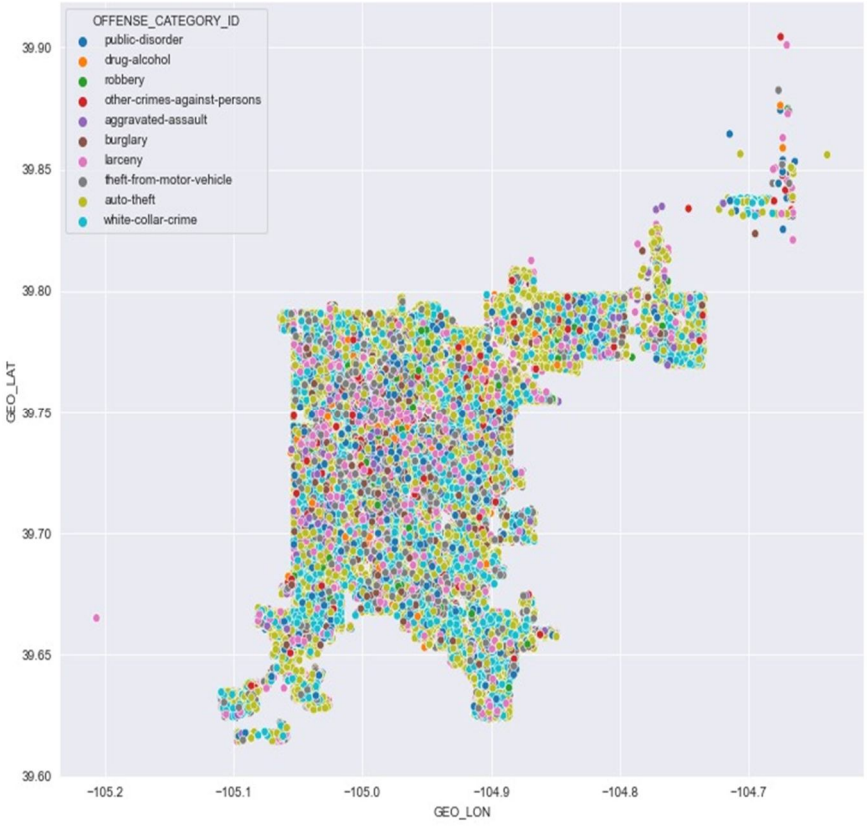


Figure 8: District-wise Distribution of Crimes

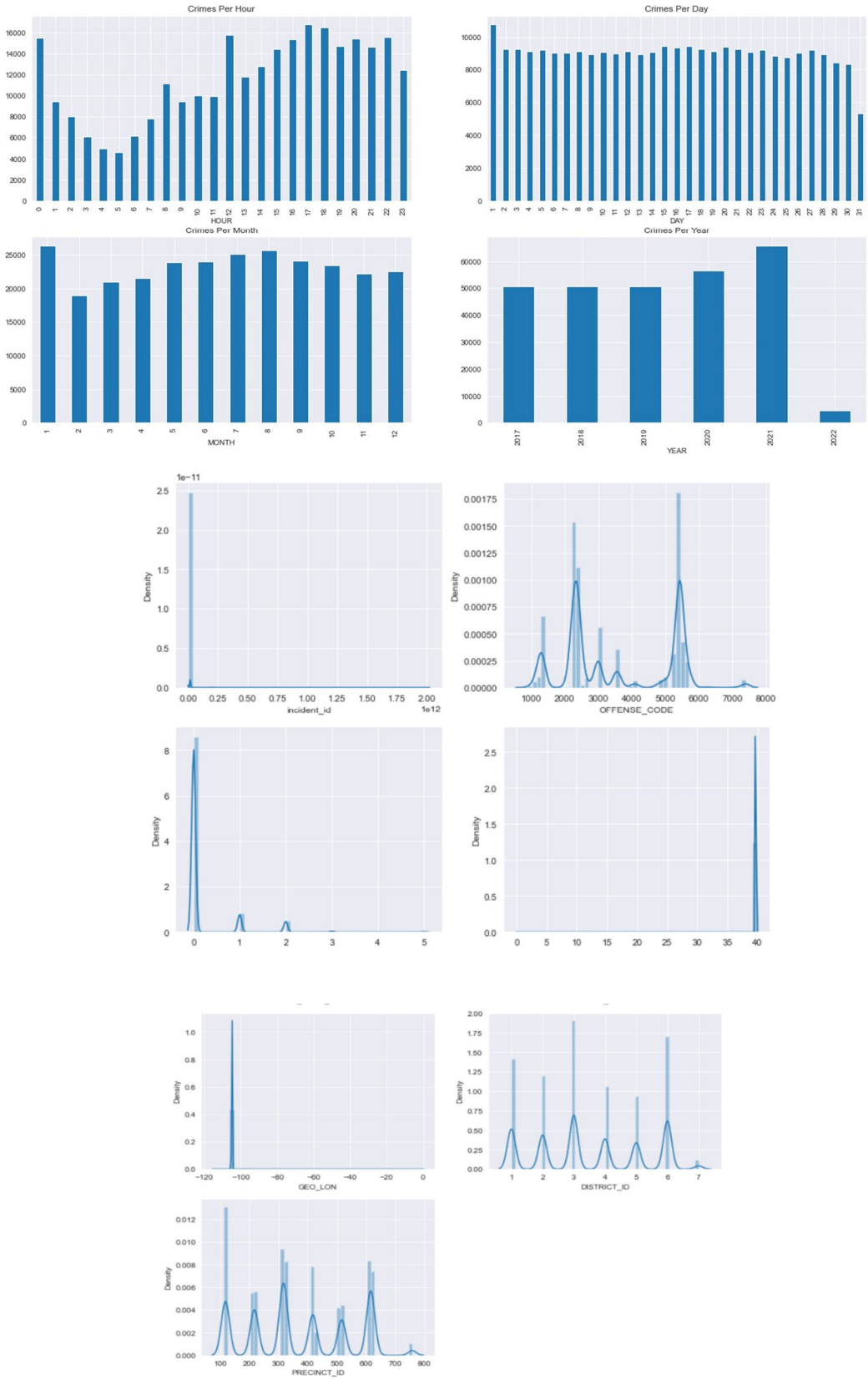


Figure 10: Distribution of Attributes

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

VI. MODEL CREATION

A. *Splitting the Dataset*

For model creation we split the data for training and testing. We defined a function that splits the dataset . The function divided the dataset in training and testing and further into the training dependent and independent variables.

```
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

features = ['GEO_LON','GEO_LAT','OFFENSE_CATEGORY_ID']

X = df[features[:2]].values
y = df[features[-1]].values
y = np.reshape(y, (df.shape[0], 1))

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
print(X_train)
```

B. *Fitting the Mode*

```
from sklearn.preprocessing import StandardScaler,MinMaxScaler,RobustScaler
#scaler = StandardScaler() #using standard scaler method
#X_train_scaled = scaler.fit_transform(X_train) #Scaling of Train dataset
#X_test_scaled = scaler.fit_transform(X_test) #Scaling of Test dataset
```

```
knn = KNeighborsClassifier(n_neighbors = 15)  #Instantiate KNN with k=3
knn.fit(X_train,y_train) #Call the fit method of KNN to train the model or to learn the parameters of model
y_pred = knn.predict(X_test) #Predict
print(confusion_matrix(y_test, y_pred))
KNN_accuracy = accuracy_score(y_test,y_pred)
print(classification_report(y_test, y_pred))
```

We fit the model using the training split of the dataset by passing the dependent and the independent variables . We passed the testing dataset in the validation dataset.

VII. EXPERIMENTAL ANALYSIS

After testing various machine learning models like Logistic Regression, Decision Tree, Random Forest, Support Vector Machine & K-Nearest Neighbor the result are as follows:

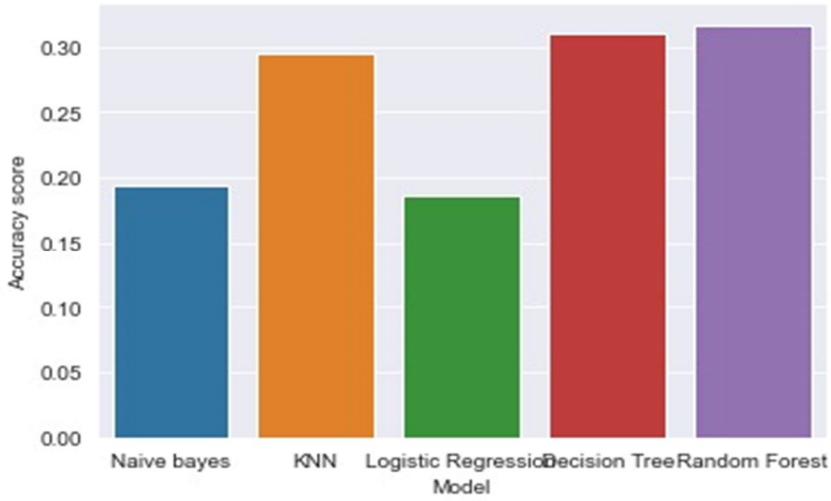


Figure 11: Accuracy of models.

We chose Decision Tree Algorithm as it produced optimal results and better accuracy.

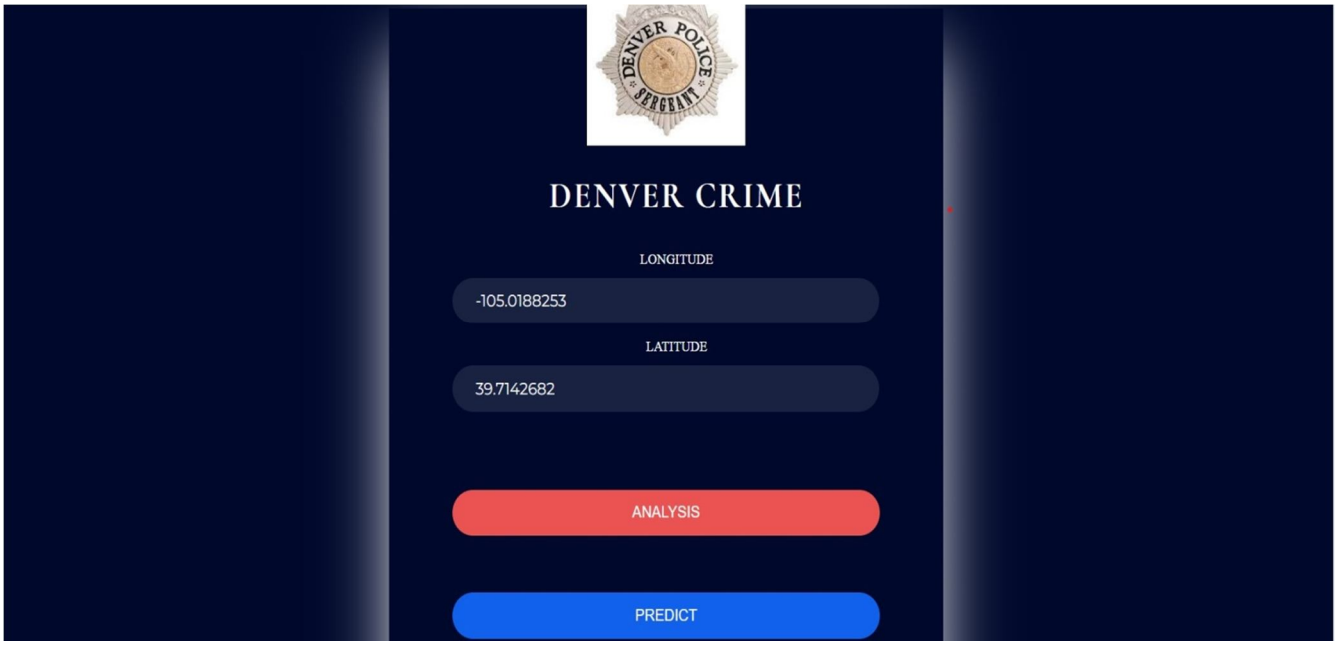


Figure 12: Web UI

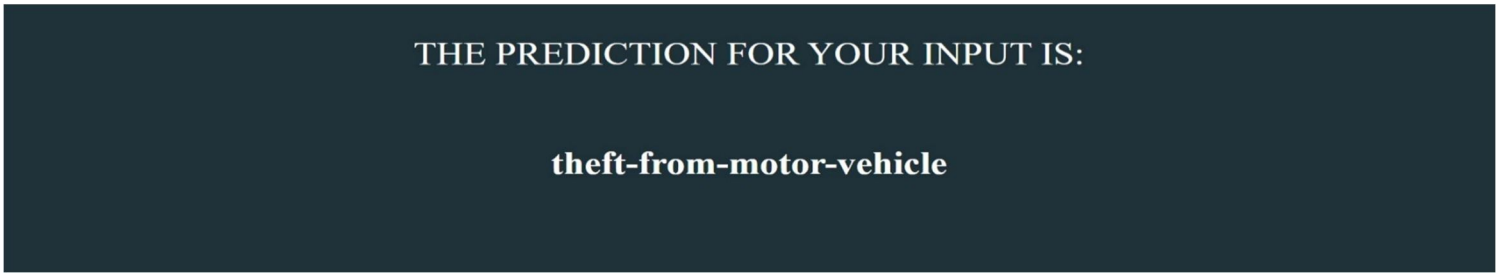


Figure 13: Prediction Outcome

After inserting the Longitude and Latitude as input it predicted which crime is likely to happen in that particular area.

VIII. CONCLUSION

Crime prediction is one the current trends in the society. Crime prediction intends to reduce crime occurrences. It does this by predicting which type of crime may occur in future. Here, analysis of crime and prediction are performed with the help of various approaches some of which are KNN, K means Clustering & Random Forest. However which model will work best is totally dependent on the dataset that is being used. 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 KNN system helps law implementing agencies for improved and exact crime analysis. The result of the optimized k-means algorithm is efficient and provides improved accuracy of the final cluster reduced the number of iterations. We know that the prediction accuracy of random forest model can be improved based on historical crime data and covariates (POI data and demographic information). From the overall prediction results, the prediction model of random forest with covariates has better performance compared with Naive Bayes model and logistic regression model. As a combinatorial classification model, random forest overcomes the limitations of single decision tree classification and can effectively avoid the problem of overfitting. Among the model evaluation indexes selected in this paper, the random forest prediction model with covariates is better than other models.

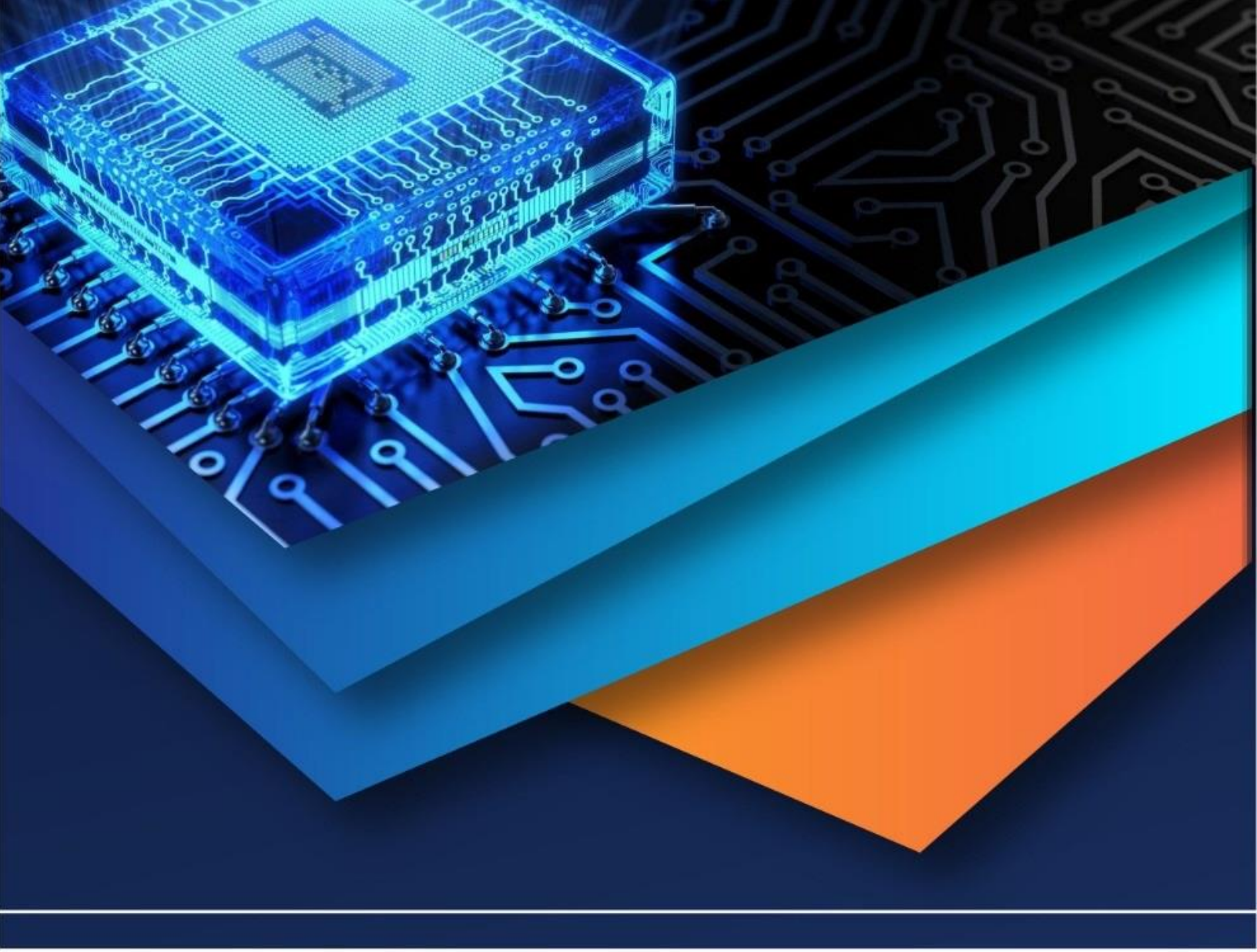
REFERENCES

[1] Pratibha, A. Gahalot, Uprant, S. Dhiman and L. Chouhan, "Crime Prediction and Analysis," 2nd International Conference on Data, Engineering and Applications (IDEA), 2020, pp. 1-6, doi: 10.1109/IDEA49133.2020.9170731.

[2] S. Yao et al., "Prediction of Crime Hotspots based on Spatial Factors of Random Forest," 2020 15th International Conference on Computer Science & Education (ICCSE), 2020, pp. 811-815, doi: 10.1109/ICCSE49874.2020.9201899.

[3] S. G. Krishnendu, P. P. Lakshmi and L. Nitha, "Crime Analysis and Prediction using Optimized K-Means Algorithm," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 915-918, doi: 10.1109/ICCMC48092.2020.ICCMC-000169.

[4] A. Kumar, A. Verma, G. Shinde, Y. Sukhdeve and N. Lal, "Crime Prediction Using K-Nearest Neighboring Algorithm," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-4, doi: 10.1109/ic-ETITE47903.2020.155.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY