CRIME ANALYSIS & PREDICTION USING DATA MINING

Pintu Prasad[1], Raj Singh[2], Rishabh Singh[3], Prof. Reena Kothari[4]

[1]Pintu Prasad, Information Technology, Shree L.R. Tiwari College of Engineering
[2]Raj Singh, Information Technology, Shree L.R. Tiwari College of Engineering
[3]Rishabh Singh, Information Technology, Shree L.R. Tiwari College of Engineering
[4]Prof. Reena Kothari, Information Technology, Shree L.R. Tiwari College of Engineering, Mumbai

---***---

***Abstract**: Crimes are a really common global problem affecting the standard of life and therefore the economic process of a society. the rise of crimes, enforcement is constant to demand advanced global information systems and new data processing techniques to enhance crime analytics and better protect their communities. Although crimes could occur everywhere, Using the concept of knowledge mining, we will analyze previously unknown, useful information from unstructured data. Predictive technique means, using analytical and predictive techniques, to spot crimes and it's just about effective in doing an equivalent. due to the increased rate over the years, we'll need to handle an enormous amount of crime data stored which might be very difficult to be analyzed manually, and also now a day's, criminals are getting technologically advanced, so there's got to use advance technologies to stay police before them. during this paper, the most focus is on the review of algorithms and predicting future crime with past analysis of knowledge.*

***Keywords**:- logistic regression, preprocessing, data collection, data visualization, analysis and prediction.*

1.INTRODUCTION:

Data Mining is that the procedure which incorporates evaluating and examining large pre-existing databases to get new information which can be essential to the organization. The extraction of latest information is predicted using the prevailing datasets. Many approaches for analysis and prediction in data processing had been performed. But, many few efforts have made within the criminology field. Many few have taken efforts for comparing the knowledge of these approaches produce. The police headquarters and other similar criminal justice agencies hold many large databases of data which may be wont to predict or analyze the criminal activity involvements within the society. The criminals also can be predicted supported the crime data. This paper presents a survey on Crime analysis and crime prediction using several data processing techniques. Our main contribution during this paper is to propose an approach supported data processing and classification method which becomes tougher when handling the large amount of knowledge, thus machine learning reduces the quantity of your time to predict crime and analysis and plot the graph of crime year wise. The organization of the paper is as follows. Section II consists of some important data processing techniques. Section III consists of an outline of the Knowledge Discovery process. The logistic regression Methods are discussed in section IV. The methods applied in crime domain are discussed in section V and therefore the paper is concluded

2. EXISTING SYSTEM:

1) Agarwal et al. used the rapid miner tool for analyzing the rates and anticipation of crime rate using different data processing techniques. Their work done is for crime analysis using the K-Means Clustering algorithm. the most objective of their crime analysis work is to extract the crime patterns, predict the crime supported the spatial distribution of existing data and detection of crime. Their analysis includes the tracking homicide crime rates from one year to subsequent.

2) Satyadevan et al. has done a piece which can display high probability for crime occurrence and may visualize crime-prone areas. rather than just that specialize in the crime occurrences, they're focusing mainly on the crime factors of every day. They used the Naïve Bayes, Logistic Regression and SVM classifiers for classification of crime patterns and crime factors of every day. Their method consists of a pattern identification phase which may identify the trends and patterns in crime using the Apriori Algorithm. The prediction of crime spots is completed with the assistance of the choice Tree algorithm which can detect the crime possible areas and their patterns.

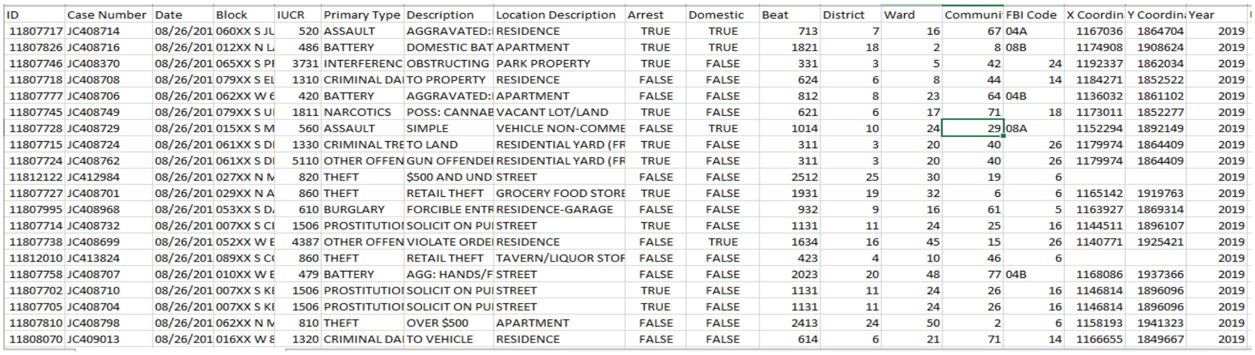
3. METHODOLOGY:

Crime analysis and prediction using data processing system have mainly five steps. the primary step is that the data collection step which we've taken from the Chicago data portal. The second step is that the data smoothing step. The third step is that the prediction step, we've used logistic regression step. The fourth step is data visualization.

3.1DATA COLLECTION:

The data utilized in this research comes from the Chicago data portal and website for open data about crime and policing in United America. This dataset reflects reported incidents of crime (except for murder where data exists for every victim) that occurred within the City Of Chicago from 2001 to present, minus the foremost recent days. Data is extracted from the Chicago local department portal.

.

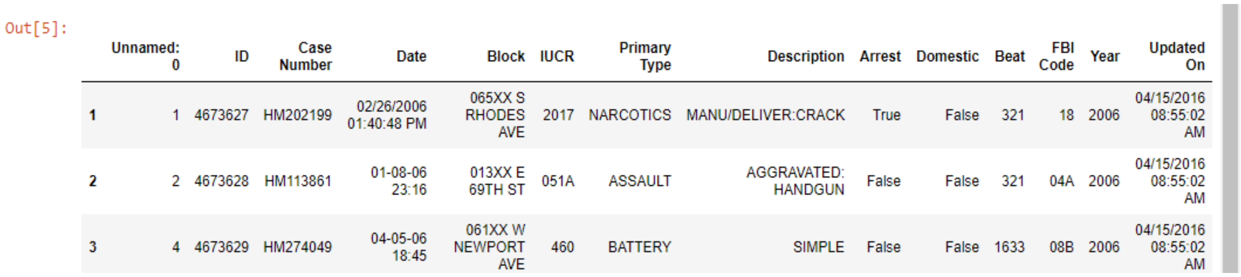


ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	Beat	District	Ward	Communi	FBI Code	X Coordin	Y Coordin	Year
11807717	JC408714	08/26/201	060XX S JU	520	ASSAULT	AGGRAVATED:I	RESIDENCE	TRUE	TRUE	713	7	16	67	04A	1167036	1864704	2019
11807826	JC408716	08/26/201	012XX N L	486	BATTERY	DOMESTIC BAT	APARTMENT	TRUE	TRUE	1821	18	2	8	08B	1174908	1908624	2019
11807746	JC408370	08/26/201	065XX S PI	3731	INTERFERENC	OBSTRUCTING	PARK PROPERTY	TRUE	FALSE	331	3	5	42	24	1192337	1862034	2019
11807718	JC408708	08/26/201	079XX S EL	1310	CRIMINAL DAI	TO PROPERTY	RESIDENCE	FALSE	FALSE	624	6	8	44	14	1184271	1852522	2019
11807777	JC408706	08/26/201	062XX W 6	420	BATTERY	AGGRAVATED:I	APARTMENT	FALSE	FALSE	812	8	23	64	04B	1136032	1861102	2019
11807745	JC408749	08/26/201	079XX S UI	1811	NARCOTICS	POSS: CANNAE	VACANT LOT/LAND	TRUE	FALSE	621	6	17	71	18	1173011	1852277	2019
11807728	JC408729	08/26/201	015XX S M	560	ASSAULT	SIMPLE	VEHICLE NON-COMME	FALSE	TRUE	1014	10	24	29	08A	1152294	1892149	2019
11807715	JC408724	08/26/201	061XX S DI	1330	CRIMINAL TRE	TO LAND	RESIDENTIAL YARD (FR	TRUE	FALSE	311	3	20	40	26	1179974	1864409	2019
11807724	JC408762	08/26/201	061XX S DI	5110	OTHER OFFEN	GUN OFFENDEI	RESIDENTIAL YARD (FR	TRUE	FALSE	311	3	20	40	26	1179974	1864409	2019
11812122	JC412984	08/26/201	027XX N M	820	THEFT	$500 AND UND	STREET	FALSE	FALSE	2512	25	30	19	6			2019
11807727	JC408701	08/26/201	029XX N A	860	THEFT	RETAIL THEFT	GROCERY FOOD STORE	TRUE	FALSE	1931	19	32	6	6	1165142	1919763	2019
11807995	JC408968	08/26/201	053XX S D	610	BURGLARY	FORCIBLE ENTR	RESIDENCE-GARAGE	FALSE	FALSE	932	9	16	61	5	1163927	1869314	2019
11807714	JC408732	08/26/201	007XX S CI	1506	PROSTITUTIOI	SOLICIT ON PUI	STREET	TRUE	FALSE	1131	11	24	25	16	1144511	1896107	2019
11807738	JC408699	08/26/201	052XX W E	4387	OTHER OFFEN	VIOLATE ORDEI	RESIDENCE	FALSE	TRUE	1634	16	45	15	26	1140771	1925421	2019
11812010	JC413824	08/26/201	089XX S C	860	THEFT	RETAIL THEFT	TAVERN/LIQUOR STOF	FALSE	FALSE	423	4	10	46	6			2019
11807758	JC408707	08/26/201	010XX W E	479	BATTERY	AGG: HANDS/F	STREET	FALSE	FALSE	2023	20	48	77	04B	1168086	1937366	2019
11807702	JC408710	08/26/201	007XX S KE	1506	PROSTITUTIOI	SOLICIT ON PUI	STREET	TRUE	FALSE	1131	11	24	26	16	1146814	1896096	2019
11807705	JC408704	08/26/201	007XX S KE	1506	PROSTITUTIOI	SOLICIT ON PUI	STREET	TRUE	FALSE	1131	11	24	26	16	1146814	1896096	2019
11807810	JC408798	08/26/201	062XX N M	810	THEFT	OVER $500	APARTMENT	FALSE	FALSE	2413	24	50	2	6	1158193	1941323	2019
11808070	JC409013	08/26/201	016XX W 8	1320	CRIMINAL DAI	TO VEHICLE	RESIDENCE	FALSE	FALSE	614	6	21	71	14	1166655	1849667	2019

Fig 1 :Data From Portel

3.2.SMOOTHING :

Smoothing may be a technique that's wont to eliminate noise from a dataset. There are many algorithm and methods to accomplish this but all have an equivalent general purpose of roughing out the sides or 'smoothing' some data. there's reason to smooth data if there's little to no small-scale structure within the data. the info danger to the present thinking is that one may skew the representation of the info enough to vary its perceived meaning, so for the sake of scientific honesty its is an important to at the very minimum explain one's reason's for employing a smoothing algorithm to their dataset. So for the smoothing purpose, we are using Dropna Method which available in pandas package. Pandas is one among those packages and makes importing and analyzing the info easier.



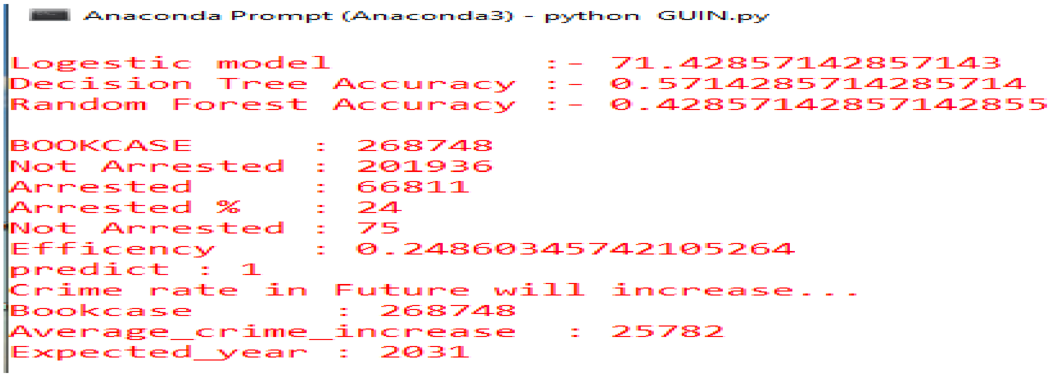
Out[5]:

	Unnamed: 0	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Arrest	Domestic	Beat	FBI Code	Year	Updated On
1	1	4673627	HM202199	02/26/2006 01:40:48 PM	065XX S RHODES AVE	2017	NARCOTICS	MANU/DELIVER:CRACK	True	False	321	18	2006	04/15/2016 08:55:02 AM
2	2	4673628	HM113861	01-08-06 23:16	013XX E 69TH ST	051A	ASSAULT	AGGRAVATED: HANDGUN	False	False	321	04A	2006	04/15/2016 08:55:02 AM
3	4	4673629	HM274049	04-05-06 18:45	061XX W NEWPORT AVE	460	BATTERY	SIMPLE	False	False	1633	08B	2006	04/15/2016 08:55:02 AM

Fig 2:: Smooth Data From Dropna Method

3.3 PREDICATION:

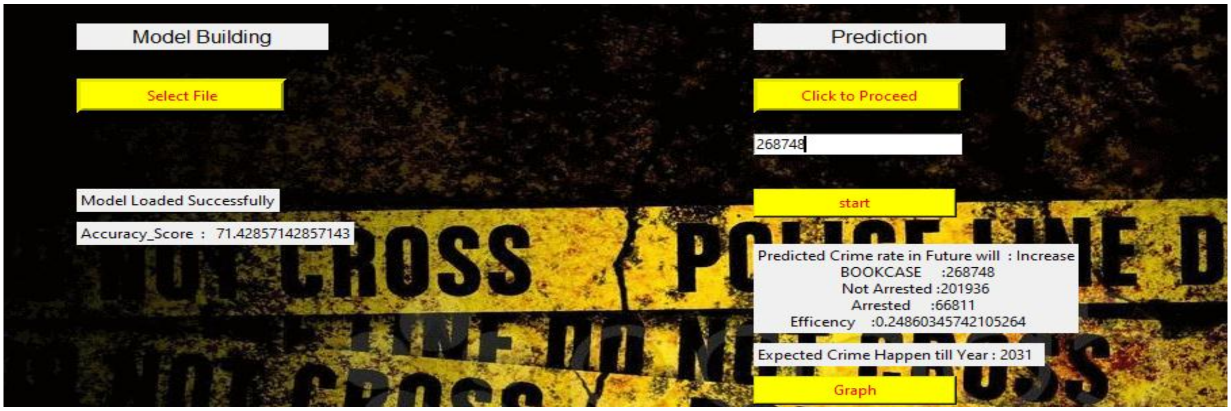
By using the various parameter for prediction we checked various algorithm like Logistic Regression Model, Decision Tree and Random Forest. Regression technique are often adapted for predication. But Logestic Regression show more accuracy on Chicago Dataset as compare to other two algorithm. Logestic analysis are used for prediction whether future crime increase or decrease supported current bookcase. In data processing independent variables attributed already known and response variables attributed what we would like to predict.Once Accuracy and Predicton is completed we've created a model by using Joblib is the quality way of serializing objects in Python. you'll use the Joblib operation to serialize your machine learning algorithms and save the serialized format to a file. Later you'll load this file to deserialize your model and use it to form new predictions.The same model type often is employed for both regression and classification. This method is predicated on the booked case entered by the user. After entering the amount they use the various parameter for prediction like Not Arrested, Arrested and Efficiency as shown in below



```
Logestic model             :- 71.42857142857143
Decision Tree Accuracy     :- 0.5714285714285714
Random Forest Accuracy     :- 0.42857142857142855

BOOKCASE       :  268748
Not Arrested   :  201936
Arrested       :  66811
Arrested %     :  24
Not Arrested   :  75
Efficency      :  0.24860345742105264
predict : 1
Crime rate in Future will increase...
Bookcase         :  268748
Average_crime_increase   :  25782
Expected_year   :  2031
```

Fig 3: Different Model Accuracy



268748

Model Loaded Successfully

Accuracy_Score : 71.42857142857143

Predicted Crime rate in Future will : Increase
BOOKCASE :268748
Not Arrested :201936
Arrested :66811
Efficency :0.24860345742105264

Expected Crime Happen till Year : 2031

Fig 4: Predication Process And Predication GUI

3.4 DATA VISUALIZATION:

Data Visualization is important for representing insight from data during a graphical manner. With an outsized amount of knowledge within the dataset, one among the best challenges is to simply communicate the hidden pattern and findings easily and understandably.to visualize the info there are many visualization techniques available.

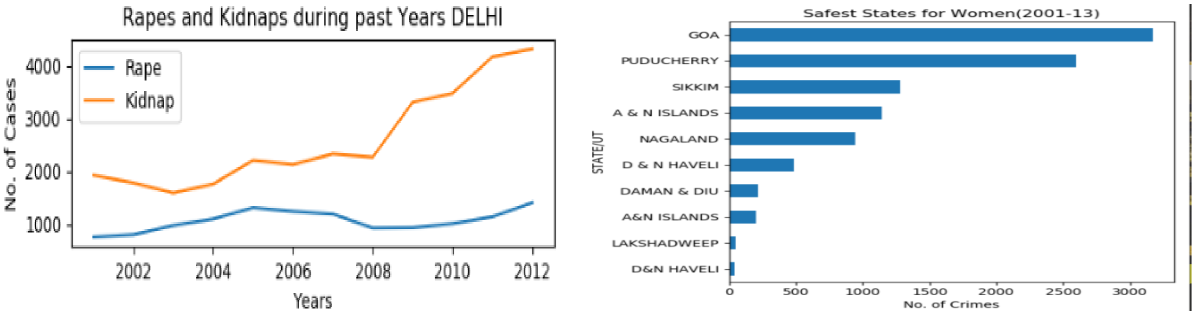


Fig5: Logistic Regression Graph And Bar Graph

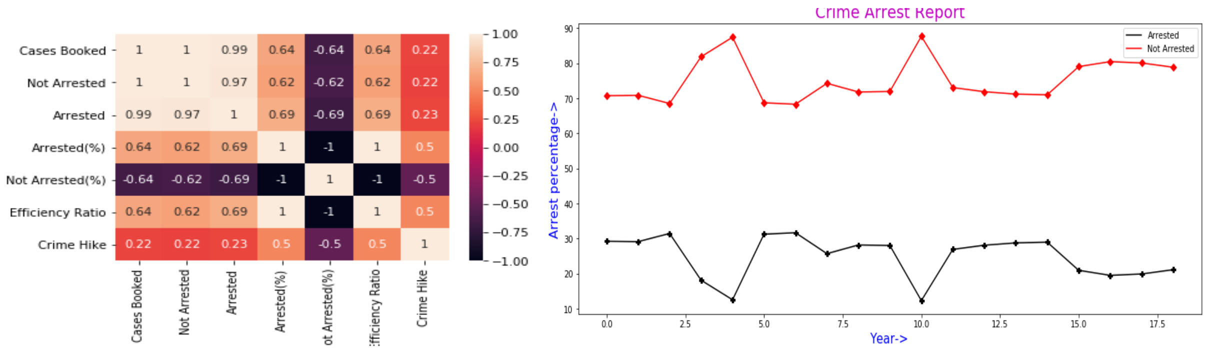


Fig6: Correlation Matrix And Basic Plotting Graph

Conclusion:

In today's world, where the typical quantity of knowledge that an individual handles has been increasing by leaps and bounds over the past few years, the use of knowledge mining techniques to extract useful information from the large amounts of data becomes important. This project mines the large amounts of data by first generating it within the sort of a dataset then preprocessing it. the varied data processing techniques, algorithms and models mentioned when applied on such datasets produce results which might be of great potential use to enforcement agencies especially. last, thus, we hope that this project performs its functions well, and works with the very best efficiency possible which it surely proves to be a boon to enforcement agencies. The functionalities of this project are often scaled up within the future. These functionalities might be Real-time data analysis of crime data: this might help us obtain crime patterns and forecasts of the longer term instantly using real-time datasets. data processing of social media to get datasets, then preprocess and analyse them to identify trends of the present crime situation during a particular place or region. Compare and display the results of all available and applicable forecasting, predicting and classification models side by side, such the user can select any of these methods.

Future work:

The proposed project mainly focuses on highlighting the crime rates around a specific state .So as per the longer term vision, another a part of the country and world are often taken into consideration. immediately it's specifically sure to a couple of crimes. Other crimes are often included further. There are certain crimes andcases which are unheard and unregistered round the globe and if they're taken into attention theaccuracy of the crime rates are often improved. Currently, the dataset consists of only registred crimes but this will be expanded to incorporate more crimes within the future. and that we will attempt to expand our project as criminal profilling

References:

[1] Alkesh Bharati1, Dr Sarvanaguru RA.K2, "Crime Prediction and Analysis Using Machine Learning". (IRJET) e-ISSN: 2395-0056 | Sep 2018| (IRJET)

[2] ChhayaChauhan1, SmritiSehgal2: A REVIEW: Crime Analysis Using Data Processing Techniques And Algorithms. ICCCA2017)

[3] Gaurav Govindaswamy1, Santhosh Kumar P3" A Survey on Crime Data Analysis Using data processing Techniques". Int. Journal of Engineering Research and Application) August 2017, pp.30-34.

4] H. Benjamin Fredrick David1 and A. Suruliandi2, "Survey On Crime Analysis And Prediction Using Data Processing Techniques". ICTACT JOURNAL ON SOFT COMPUTING, APRIL 2017,

[5] Sunil Yadav, Meet Timbadia,." Crime Pattern Detection, Analysis & Prediction" ICECA 2017