University of Sindh Journal of Information and Communication Technology (USJICT)
Volume 3, Issue 1, January 2019
ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

PREDICTIVE POLICING: A Machine Learning Approach to Predict and Control Crimes in Metropolitan Cities

Jibran R. Khan, Farhan A. Siddiqui, Nadeem Mahmood, Muhammad Saeed, Qamar ul Arifeen

Department of Computer Science – UBIT, University of Karachi
jibran_rasheed@hotmail.com, farhan@uok.edu.pk, nmahmood@uok.edu.pk, saeed@uok.edu.pk, arifeen@uok.edu.pk

Abstract: Security is the one of the basic need of human's life and biggest challenge of history that cannot be diminished at least in metropolitan cities like Karachi. It can only be controlled by efficient resources allocation and effective strategies with forthcoming insight of criminal moves. Big data analytics with the support of machine learning algorithms makes it possible to deals with huge amount of data, extract hidden inter-connection, pattern and meaningful information. This paper, proposed the model for the predictive policing system and built test model using k-means and naïve Bayes methodologies for street crime in Karachi region. The model is then run under R and WEKA environment which produced accuracy around 70%.

Keywords: Predictive Policing; Machine Learning; Crime Prediction; Clustering; K-Means; Naïve Bayesian.

I. INTRODUCTION

Security, is one of the biggest challenge of history, irrespective of domain that it cannot be eliminated but can only be minimized by knowing what will happen in near future, to ensure basic human's life safety needs. It is the only factor which consistently increases by time and with evolution of technology. The rapid growing scale of crime make it hard to travel, run business and amuse life with family. Reflecting many serious incidents of Pakistan such as Peshawar APS attack, Kamra base attack and Karsaz base attack, Safora bus attack, and many other majors. Criminal activities are technologically erudite which became hard for official personnel to defend and untie the complexities of crime in mean time[1], [2].

Criminology is the scientific study of understanding criminal behavior, crime nature, and planning anti-crime strategies by identifying the characteristics of crime. Crime analysis is one of the core activity of criminology which deals with the exploratory of crime incident, identifying crime patterns and their links with other terrifying incidents and criminal groups. In the presence of huge amount of data, complexity and frequent committing crime rate, it almost impossible for security authorizes and their personnel to manually study and unfold the hidden secrets, inherited complexities[1], [3], [4]. Thus, the task is distributed over several officials which leads to leftover relevant information and might could not cover all aspect of crime in a mean time. Since long time, statisticians and criminologists professional facilitating by applying their skills to predict next incident with some degree of success, but these efforts are not to the mark and every failure putting strain on the exiting methodologies[5].

In the past few years, state of art data science applications technologically appeared as promising reliable solution for business strategies, smart decision making, recommendation system, marketing, trending and in determining weather and environmental likelihood[6][7]. Inopportunely, where people are more concern about security to run their business and live fearless life, are expressing less interest in the modern predictive policing approach. Based on facts that 1) every incident is not just a single entry of record that happened but holds valuable information that could serve in mitigating crimes. It is said that past is the true providence of future[8], because the surrounding parameters of incident scene like society, environment, type etc, encourage criminal elements and attracts their point of interest in committing crime easily which describe the probability and comfort of conducting dreadful activities. 2) Crime is predicable[9], it is a human behavior or nature of people of doing things. Psychologically is has been proved that the human nature is irreversible, which cannot be completely changed or shaped into new one, but can be confined by inferring monitoring and surrounding limitations[7]. Human behavior or nature periodically evokes itself which repeats acts in feasible environment and hyperactive or addicted nature appeared accidently irrespective of its environment. 3) Human reasoning gets fail with the proportionality of involving amount of data. Thus, it is almost impossible for human intelligence to tackle multiple records, dealing with numerous working files, remember all the historical linkages of all crimes or criminals, map them and extract their relation in a time bounded circumstances.

Inspired by modern criminology study and technological advancement that support dealing with huge amount of data with high computational power and predicting future hot spots. This paper presented system design for predictive policing to counterfeit street crime such as mobile theft and

snatching along with the simulated results of machine learn model using different algorithm approaches and their performance comparison. The remaining paper is divided as Section 2 presents some studies of rising interest in predictive across the world[10]. Section 3 discusses the data set, its source, nature, and their prerequisite steps. Section 4 tells about the system design and its working environment. Section 5 confer on various algorithm used in building predictive models. Section 6 illustrates the predictive outcomes and their performance comparison. Then finally, Section 7 close the discussion by presenting conclusion of work.

II. LITERATURE REVIEW

This section presents the world's review of adoption of predictive policing approach, which is either practically implemented to overcome crime ratio in the society and various methodologies that can be useful in adopting such system.

In [11]M. Camacho-Collados et.al proposed the DSS as predictive policing opportunity for Spanish National Police Corps (SNPC). SNPC collaboration and the need of preventing crimes aids in determining definite problem space and getting relevant data. The mathematical model is used to identify the district region for better distribution and predictive techniques utilizes for future forecasting. The trained model is then tested in Central District of Madrid, where predictive results majorly accurately forecast the future incidents and DSS smartly suggest resource distribution in the targeted region. Another study found that is implemented for the Korean Police Agency, where crime rate exponentially rising and life become miserable especially for women and children. A. Nasridinov et.al [12] work was the first implementation of predictive policing in South Korea, on real crime dataset gathered from different sources. He used various profound machine learning algorithms such as K-nn, SVM, Neural Network (NN), DT and others. Furthermore, authors discuss the impact of unstructured and variant dataset on the predictive system performance. Same algorithm on viable dataset showed the variance in model prediction also system performance. Besides this fact, on contrary, the proposed system promisingly helps in reducing the crime in the said region and significantly predict accurate crime instances. Similarly in[13], C.H Yu et.al presented new feature selection approach by developing Cluster-Confidence-Rate-Boosting (CCRBoost) algorithm to grouping the crimes in the spatio-pattern on the real world dataset provided by the US police department of a northeastern city. City dataset covered 90 sq. miles region which is distributed over 800 meters grid cell to point the area. Whereas, built system 80% accurately locate the crime scenes and potentially assist police department in decreasing crime in city.

Besides the above few studies, study also covers the various others, some of them are summarized in the following but not limited to Table I.

TABLE I. SUMMARY OF LITERATURE SURVEY

Ref	Studies	Year	Type
[8], [14]–[17]	5	2002-2005	Discuss different methodologies, possibilities and factors and factors that influencing crime and helps in predicting them.
[18]–[21]	4	2007-2011	Reviewing the techniques, discuss the issues, and stating steps in determining predictive policing.
[5], [22]–[24]	4	2012	Enhancing and contributing to existing approaches supporting ideas in Crime investigation and prediction using machine learning.
[3], [9], [25]–[29]	7	2013	Attracting world attention reflecting the needs of advance policing strategies and foresight capabilities.
[7], [12],[13], [30]–[37]	11	2014	Many works research and pilot works supported by local governments. Promoting predictive policing. Performance evaluation, new methodologies, reviews.
[6], [11], [38]–[48]	13	2015	Different techniques comparison, risk assessment of modern system, discussion on choice of DSS better for short sight or long term. Spatial, hot spot GIS etc.
[1], [10], [49]–[57]	11	2016	ML adoption by organizations and agencies to integrate with their system. Agent based modeling, pattern trailing, assumption and evaluation of applications. Social network based crime analysis.
[2], [58]–[65]	9	2017	Big data, BI, AI, social media and human computer interaction role in predicting crime and trends in cyber and physical world security. Also, concern with its affordability.

The above summarization not only captured the academics and researcher's insight for future policing system but also the prestigious organization and authorities require such solution to avoid any mishap such as National Institute of Justice, Korean Police Agency and Seoul Data Agency[12], National Institute for Justice (NIJ) [21] and RAND[9] USA, Justice and the RCMP of Canada[8], Spanish National Police Corps (SNPC)[11], MPS Territorial Policing BOCU London[44] and many others.

III. DATA AND METHODOLOGY

Various techniques and methodologies can be used to achieve the above mentioned objective, the study and work region is Karachi city, which is widely spread, highly populated and having nested complex road network. In order to investigate the said goals in relevant manner, actual & complete crime data is required and analyzed accordingly.

A. Data & Sources

For real world dataset, is said that it is never be as you expect. What you get is usually what you not require. Real datasets mostly unstructured, noisy, inconsistent incomplete and consisting missing data in records. It is mainly due to the lack of standardization, viable requirements and need of organizations. Thus, this trouble in data processing and

knowledge discovery. This improper and imbalance dataset entail data cleaning to improve the data quality and make it more meaningful in information extraction. Data pre-processing is not an easy task which is actually consume around 70% of efforts of complete data analysis and predictive development[66].

Abating these issues, the dataset and other resources that are utilized that should render the factual ground knowledge of crimes in city. The expressive crime map presentation, analysing and producing results requires the database/dataset which should at least contain the date, time, area, subarea, location base information of specific crime (GEO coordinates). In addition to this town based information, UC based information, individual police station and its jurisdictional information, and arrested or not. All these information is also necessary for better analysis, portraying crime pattern, criminal nature, groups, interlinks and their associative region. The possible attainment of such desirable data can be obtained by different sources such as CPLC, security agencies, government organization, and Karachi Police.

B. Data Collection

The hypothetical data is useful in understanding and structuring that made the system closest for portraying the real picture. The time, area, geo-location cords information are considered as the most important attributes for developing the system and mapping crimes. The google maps is used to collect the GEO coordinates with error approximate of 20 meters, as it plays and important role in location based monitoring and mapping crimes.

C. Availability and Accuracy of Data

In the predictive analysis, data is the key essential and important element of system. Most often the data required for modelling are obtained from databases or flat files which is full of errors. While, the most predictive models require clean formatted data[67], and its credibility is prerequisite condition of any such product. Whereas, the data selection is mainly based on the project nature and its ultimate goals to achieved[64].

The dataset used in project is of mobile crime contains the data from all over the Pakistan, which has various attributes like case id, mobile brand, mobile model, crime type, area and landmark name, date and some un-useful and unnecessary attributes.

While the provided dataset is not up to the mark as needed and also missing the critical attributes like time, location code, police station jurisdiction information, town information and union council base information. In addition to data accuracy there are lots of multiple spell error, improper and irrelevant area and landmark names are found.

D. Dataset Attributes

The given dataset of Karachi city is used for building the predictive model, analysis and tune system. The dataset consist of 16 attributes shown in Table II, and more than six lacs records from the period year 2005 to 2013.

We distinguish two basic roles a variable:

- Independent (also called predictor, explanatory, feature) – these variables describe the properties of objects which we want to use as the basis for making inferences.
- Dependent (also called response, explained, target) – these variables describe the features of the object which we want to make inferences about.

TABLE II. ORIGINAL ACQUIRED DATASET

Attribute	Type/Description	Attribute	Type/Description
COMPNO	Int/Complain number	INCDATE	Dt/Incident date
COMPDATE	Dt/Complain date	AREANAME	Txt/Incident area name
IMEINUMBER	Int/Mobile IMEI number	INCLANDMRK	Txt/Incident location name
SIMNUMBER	Int/Mobile SIM number	PSNAME	Txt/Police station name
BRANDNAME	Txt/Mobile brand name	COMPNAME	Txt/Complainer name
MODELNAME	Int/Mobile model number	CNICNUMBER	Int/Complainer CNIC number
COLNAME	Txt/Mobile color	CADDRESS	Txt/Complainer address
CRNAME	Txt/Crime type	CHOMEPHONE	Txt/Complainer contact number

.

IV. DATA PRE-PROCESSSING

The pre-processing data comprises of several tasks which include cleaning of inconsistent and noise in information, data integrating process, transforming the dataset into intended and machine process-able form and save it into data warehouse or storage location[66]. Furthermore, in addition to this, adding, splitting, merging, and extracting necessary and hidden data attributes are also done in this step as shown in the Fig.1. The resultant dataset is then feed into predictive model that is processed by the machine learning algorithm for training and testing[68], [69].

Basic exploratory observation found that the dataset have missing values and number of inoperable implausible attributes. Among the above mentioned attributes, this work only focusses on potential crime related fields. Because, the performance efficiency and quality of knowledge discover process in machine learning application is directly proportional and dependent on the pre-processed or fed data quality[66]. Following are a number steps taken to clean the dataset and illustrated in Fig.2 and transformed dataset is shown in Table III.

1. Import appropriate data into environment.
2. Explore dataset to determine data health.
3. Select data range or records to process.
4. Identify potential and unusable attributes.
5. Replace and fill missing information of record(s) with default, mean, model derived or global constant value(s).
6. Remove incomplete information record(s), if step 6 not fill the gap(s).

7. Remove the complete record(s) contain(s) garbage data.
8. Remove redundant and duplicate record(s) based on attributes tuples (INCDATE, COMPDATE, COMPNAME, IMEINUMBER). The given dataset is merge of various sources, this tuple helps in identifying duplicate or redundant entry in dataset.
9. Remove unusable attribute(s): COMPNO, PSNAME, SIMNUMBER, IMEINUMBER, SIMNUMBER, MODELNAME, CNICNUMBER, COLNAME, CADDRESS, and CHOMEPHONE.
10. Split attribute(s) into new potential attribute(s): INCDATE fragmented into day, month, year, weekday name i.e Tuesday, week number and reported time.
11. Add appropriate or extracted attribute(s): Add Town and incident location geo-coordinates with help of Google API and named them as MALong, MALat for town longitude and latitude, SALong, SALat for incident location.
12. Grouping attribute(s) values for more meaningful information: Time into Time four (4) categorical time slap window 12am-5:59am, 06am-11:59am, 12pm-5:59pm, 06pm-11:59pm.
13. Removing multi-spell error by renaming values to standard spell as used by Google which helps in finding coordinate and also will be helpful to plot the final results on map. Instead of listing all taking a single error to elaborate working i.e Gulshan-e-Iqbal, gulshan Iqbal, gulshan e Iqbal, gulsahneiqbal gulshen iqbal to Gulshan-e-Iqbal.
14. Save transformed dataset as clean Dataset.

TABLE III. PROCESSED TRANSFORMED DATASET

Attributes	Description	Type
Brand	It shows Mobile Brands, which consist of 8 distinct mobile brands; Black Barry, HTC, iPhone, LG, Megagate, Motorola, Nokia, Samsung, and Sony are numbered as 1 to 9 in alphabetical order.	Numeric
Crime	Crime Type focusing on only two street crimes; Snatch and Theft	Numeric
Date	Incident Date. Calendar date	Date
Day	Incident day of week in number, where Monday is day 1.	Numeric
Week	Week of the year in number i.e. 27, 28, 29….	Numeric
Month	Denotes the month of the year, where January is 1 and December is 12	Numeric
Area	Incident area (main area) name	Text
Landmark	Incident sub-area (landmark, street, etc.) name	Text
MALong	Main area Latitude	Numeric
MALat	Main Area Longitude	Numeric
SALong	Sub-area Latitude	Numeric
SALat	Sub-area Longitude	Numeric

V. TECHNIQUE

Exploratory Data Analysis (EDA) is a heart of predictive system that was named by John Tukey, use for analyzing and summarizing dataset by their characteristics with visual techniques[70]. This approach does not necessary require any statistical model but can be used for knowing what data reveals without hypothetical and formal model testing[69]. EDA visual techniques are generally very simple in nature and quite expressive that easy to understand. These consist of:

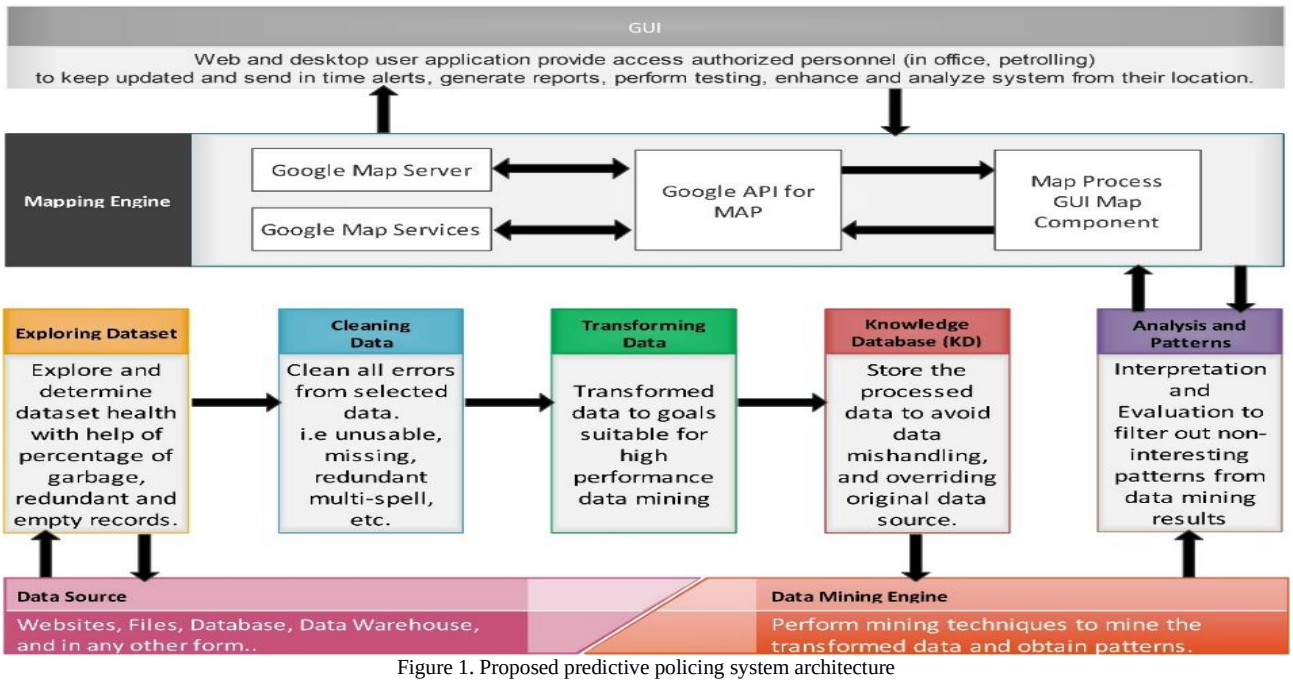

Figure 1. Proposed predictive policing system architecture

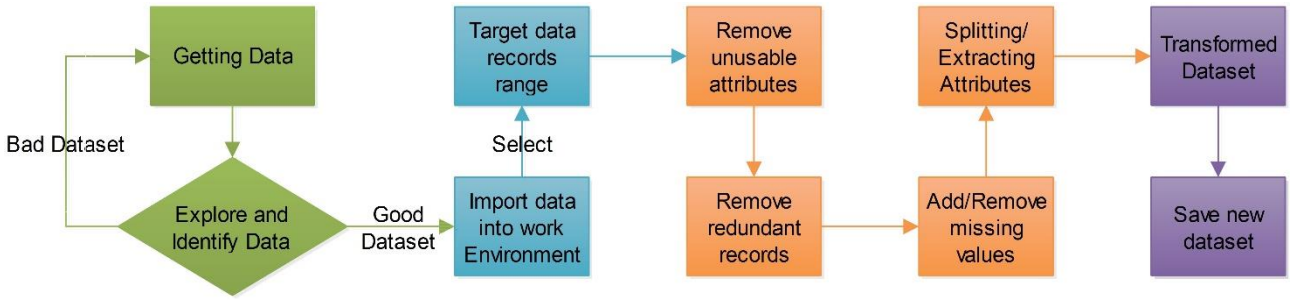

Figure 2. Data cleaning process flow

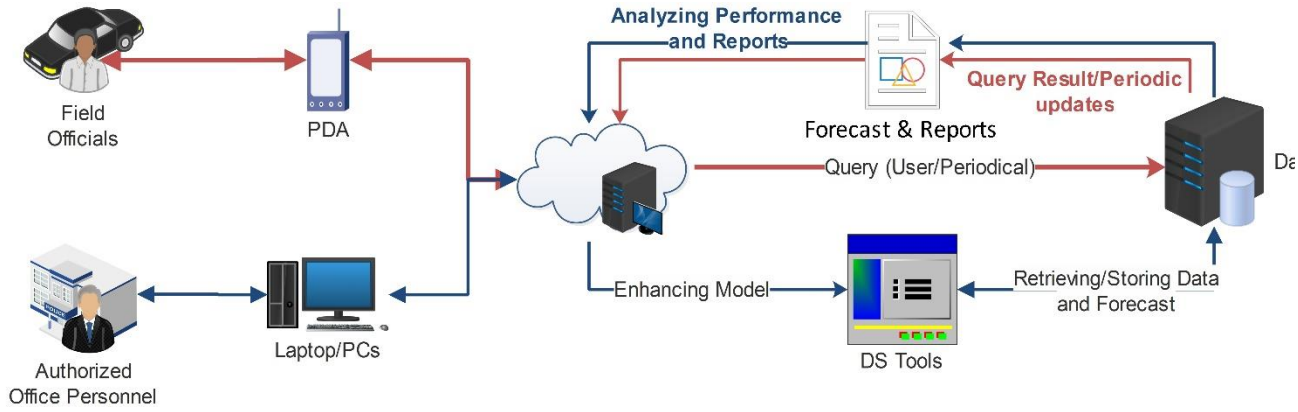

Figure 3. Predictive policing system process model

1. Raw data graphs such as histograms plot, box plot, and probability plots.
2. Statistical graphs such as standard deviation, box, mean plots and others.
3. Positioning graphs such as max, min, pattern-recognition, multiple plots and many more.

VI. System Model

The underlying system flow design serve clients to process, store and retrieve the data shown in Fig. 3. According to the present knowledge & research study, presented the design is fully optimized and flexible to for future enhancement and adjustment. New factors can be easily added to the system to process dynamically in order to improve the effectiveness and efficiency of system.

VII. Algorithms and Mining Technique

Since the idea enlighten the possibilities of making machines intelligent to use their advance and high computational power assist-able in better decision making, early alarming time sensitive applications. People have been proposed numerous machine learning algorithms that are designated to deal with their respective problem approach and broadly categorized them into two subject; Clustering and Classification.

A. K-means

Among the other machine learning algorithms, K-means[23], [34] is one of the simplest and less complex clustering algorithm[38] which comes under unsupervised learning technique[57]. Clustering mechanisms are mainly use for partitioning the data into their respective heads based on the characteristics similarities[31], [63], it does not predict future but sorting data into partitions. This helps in cluster analysis to learn the behaviour of corresponding entity to identify which geo-spatial region it belongs to[1]. In the following presents flavoured the K-means algorithm used in this work for sorting dataset.

1. Setting number of K cluster as number of Main area in dataset. i.e 93.
2. Sorting dataset w.r.t to main area, centroid is determine by the Longitude and Latitude of main area.
3. Nearness Proximity is determined by Euclidean distance, cosine similarity, Haversine formula to estimate distance and also will help in mapping crime.
4. Grouping the sub-area under their respective main areas.
5. Repeats the iteration unless centroid stop changing
6. End if all are stable and partitioned.
7. Save the processed dataset.

B. Naïve Bayesian

Unlike K-means algorithm, Naïve Bayes[21] is one of the supervised and well-known classification machine learning approach use for predicting the future instances[22]. It has been extensively utilized in various studies and research works produced surprising result in innumerable domain[6] which is known for its better performance and accuracy as compare with others[52]. It uses Bayes theorem for computing the probability of every class from underlying evidence[1]. The general naïve Bayes equation used in building predictive model is shown in (1)

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

$$P(c|x) = P(x_1|c) \times P(x_2|c) \times P(x_3|c) \times \dots P(x_n|c) \times P(c)$$

$$\mathrm{P}_c = \frac{\mathrm{P}(crime|main\ area, sub\ area,\ date, day)}{P(main\ area, sub\ area, date, day)}$$

$$P_c = P(main\_area) \times P(cirme|main\_area) \times P(sub\_aera) \times P(crime|sub\_area) \times P(date) \times P(crime|date) \times P(day) \times P(crime|day)$$

Where P(c|x) donates the posterior probability, P(x|c) is likelihood, P(x) is predictor probability and P(c) is class probability.

VIII. RESULTS AND DISCUSSION

This section presents the few core results of study that acquired by using K-means and Naïve Bayesian algorithms for clustering and predicting crimes in city. Fig. 4 illustrates the crime clustering based on the density. Among the 93 distinct main areas, density clusters results prominence 6 major area that are similar in nature which can be further distributed intro two types. The upper clusters zone consisting Gulshan-e-Iqbal, Gulistan-e-Jouhar, Nazmabad and lower clusters zone comprises of Defence, Korangi and Sadar areas. While, if we distribute the resultant clusters based on their type as shown in table we found an interesting fact that Gulshan-e-Iqbal, Gulistani-e-Jouhar and Defence are the areas belongs to upper middle class and upper class areas people. Whereas, Nazmabad, Sadar and Korangi are few the busiest business and industrial areas of city that have high traffic and activity areas. Table IV, shows the top 10 crime areas of city among 93 areas. From the Table IV, Sadar, Gulshan-e-Iqbal and Gulistan-e-Juhar have the highest crime intensity and is the busiest area of city. Where, educational institution and offices are mainly situated in Gulshan-e-Iqbal and Gulistan-e-Juhar, while central trading market and head offices large enterprises are situated in Sadar town. Gulshan-e-Iqbal and Gulistan-e-Juhar are situated adjacent to each other with wide running road and clear strret while Sadar Town located opposite to them with congested people traffic and comparatively less wide streets. But, both (Gulshan-e-Iqbal, Gulistan-e- Juhar town and Sadar Twon) are similar in their nature of people and vehicle traffic, visit routine and working .

TABLE IV. TOP 10 CRIME AREAS

Area Name	July	August	September
Gulshan-e-Iqbal	40	52	28
Gulistan-e-Juhar	24	23	20
Saddar	21	21	23
Clifton	15	12	4
Malir	14	6	5
Defence	12	18	3
North Nazimabad	12	15	10
Shahra-e-Faisal	11	8	8
Nazimabad	10	11	6
Liaquatabad	9	10	8

TABLE V. WEEKLY CRIME INTENSITY BY MOBILE BRAND

Week	BLACK BERRY	HTC	IPHONE	LG	MEGAGATE	MOTOROLA	NOKIA	SAMSUNG	SONY	Other Brands Total Hits	Nokia Total Hits	Ratio Other vs Nokia
07-02	2	3	1	1	0	0	47	4	0	11	47	1:8
07-09	2	3	4	1	0	0	65	10	2	22	65	2:7
07-16	2	3	1	0	1	0	68	3	1	11	68	1:8
07-23	2	8	4	1	0	0	61	3	2	20	61	2:7
07-30	6	2	2	2	1	0	66	10	0	23	66	2:7
08-06	2	3	3	1	0	0	63	7	1	17	63	2:7
08-13	5	2	3	0	0	0	63	12	0	22	63	2:7
08-20	6	1	4	0	0	0	64	10	2	23	64	2:7
08-27	3	2	3	0	0	0	66	6	2	16	66	2:7
09-03	0	4	3	0	0	0	67	6	1	14	67	1:8
09-10	4	2	1	0	0	0	61	4	1	12	61	1:8
09-17	1	5	2	0	0	1	52	5	2	16	52	1:8
09-24	1	1	0	1	0	0	46	1	0	4	46	1:9

The following Table V illustrates the weekly crime rate based on mobile brand. This reveals an interesting fact that among all brands Nokia is the only and highest lost rate or targeted brand. This also state that Nokia is the quit affordable and inexpensive brand among all or in other words based on city situation people more likely to buy the Nokia instead of any other brand which due to its affordability (low price), low loss cost, availability on every next mobile shop that makes it highly demanding mobile brand in the situation at that time. The probability of hitting other bands with compare to Nokia is about 1:8 (Other: Nokia). In every 8 hits there might be chance to get the high end mobile, or in another way, only one out of eight (1:8) people is prefer to use high-end or other mobile brands instead of Nokia.

Table VI presents the weekly crime rate and crime distribution by type i.e: Snatch and Theft. The statistics of Table 6 and recalling the above mentioned parameters of top 3 highly crime areas in our analysis i.e: Gulshan-e-Iqbal, Gulistan-e- Juhar town and Sadar Twon. We can conclude that it is comparatively more comfortable and easy for criminals to snatch your mobile or lost mobile while people are in high congested market area or walking alongside on wide running road, instead of someone silently pickup your pocket. From the Dataset, we found that the majority of theft attempts are happened in market areas i.e. Sadar town (is one of market areas) while snatching incidents are majorly reported in other areas which covers institutional areas, private company or offices and restaurants situated in commercial lane around residential areas i.e Gulshan-e-Iqbal, Gulistan-e- Juhar

TABLE VI. WEEKLY CRIME RATE BY CRIME TYPE

Week	Snatch	Theft	Freq.
2012-07-02	24	34	58
2012-07-09	52	35	87
2012-07-16	38	41	79
2012-07-23	43	38	81
2012-07-30	52	37	89
2012-08-06	40	40	80
2012-08-13	48	37	85
2012-08-20	57	30	87
2012-08-27	52	30	82
2012-09-03	44	37	81
2012-09-10	41	32	73
2012-09-17	38	30	68
2012-09-24	30	20	50

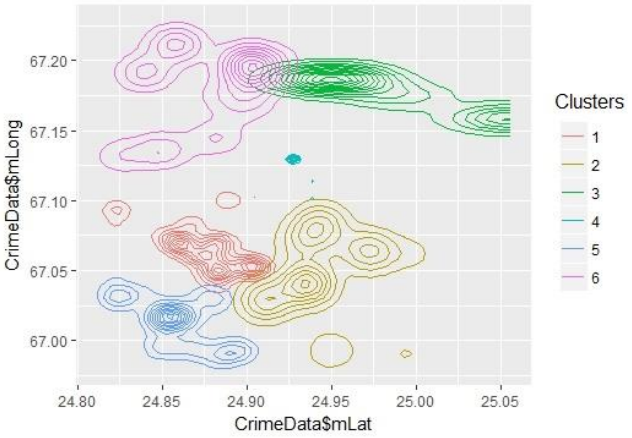

Figure 4. Crime density clusters

Table VII shows predictive model performance summary. Model is trained on different scale ratio in different cycle to test and identify the most suitable point to splitting dataset with the window of 5% from 60% to 90% dataset, and the remaining was used for testing. Experiments found that the optimum result of model using NB gained at 80-20 ratio. Above the 84% or below 80% the accuracy of model gradually decreasing and the min 67.8% accuracy recorded. Minimum 5 related features are require to find out the expected next hit. These potential features are month, week, day, time, and crime type. Table VIII shows the prediction accuracy of model in which we are identifying what type of crime could be happen in future. We successfully achieved significant accuracy of about 83%.

TABLE VII. PREDICTION SUMMARY

Correctly Classified Instances	83.2%
Incorrectly Classified Instances	16.8%
Kappa statistic	0.6565
Mean absolute error	0.253
Root mean squared error	0.3474
Relative absolute error	51.3201 %
Root relative squared error	69.9615 %
Total Number of Instances	10000

TABLE VIII. PREDICTION ACCURACY BY CLASS

	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Snatch	0.88	0.229	0.83	0.88	0.854	0.91
Theft	0.771	0.12	0.835	0.771	0.802	0.91
Avg.	0.832	0.181	0.832	0.832	0.831	0.91

Based on the given dataset (July, 12 – Sep-12), Fig. 5 depicts the next day hit. It is clearly seen that potential crime scene would be Gulshan-e-Iqbal, University road, and the Gulistan-e-Jouhar area are under threat. From the cluster analysis we know that these areas have more crime density and other crime instances (near Frere and Nasir colony) which has very low probability of occurrence.

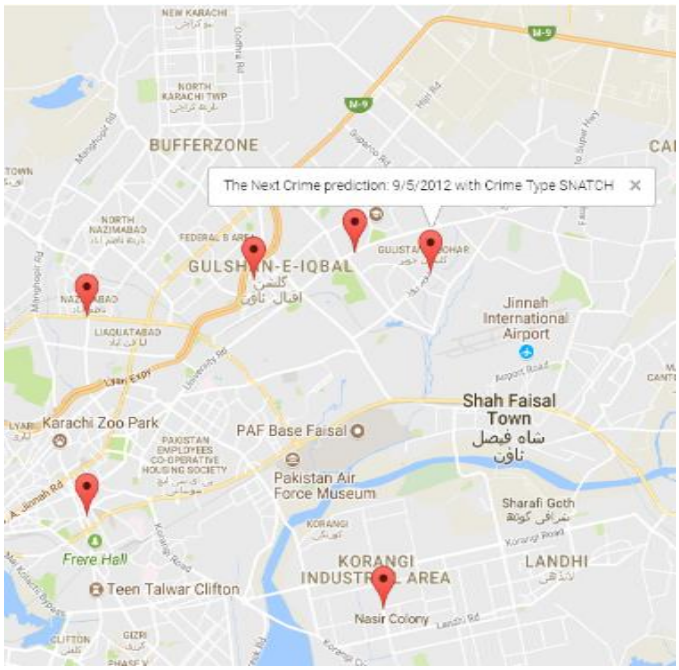

Figure 5. Next day predicted crime result

IX. CONCLUSION

Inspired by the modern criminology study and possibility of future projection by advance technology, prevalent mishaps of terrifying events, uncertainties and making sure the human life safety and enriching the law enforcement strategies. Paper discuss the worth of predictive policing, advantages and its frequent adoption by various countries that shows prolific results in a small period. This paper proposed crime analyst and predictor (CAP) for identifying and weaken the forthcoming criminal actives. The initial proposed system is built on most sophisticated approaches of machine learning; K-means clustering and Naïve Bayesian classification that are less complex outperform as compare with other techniques. Paper deals with the unstructured and noisy Karachi street crime dataset that holds mobile snatch and theft records. Foremost, data passed through the cleaning process, transformed it into new data and then forwarded into built model as input. The model, first partitioned the data into clusters using K-means and then perform the prediction using naïve Bayesian which predict the crime date and location as output. Simulation is performed under R and Weka environment, in which result shows the accuracy around 70% which is much motivating and promising in final real time predictive application. In future work, with the local government authorities and industrial partner O'Reference liaison will developed the proposed system to shape the idea into productive real-time application to maximize the efficient scare resource allocation.

X. ACKNOWLEDGE

We like to be very thankful to CPLC for cooperation and providing crime data. No other government and private organization and agencies support or funded this research work.

REFERENCES

[1] H. Hassani, X. Huang, E. S. Silva, and M. Ghodsi, "A review of data mining applications in crime," Stat. Anal. Data Min. ASA Data Sci. J., vol. 9, no. 3, pp. 139–154, 2016.

[2] E. Ahishakiye, I. Niyonzima, and R. Wario, "A Performance Analysis of Business Intelligence Techniques on Crime Prediction," Int. J. Comput. Inf. Technol., vol. 6, no. 2, pp. 84–90, 2017.

[3] S. Shojaee, A. Mustapha, F. Sidi, and M. A. Jabar, "A Study on Classification Learning Algorithms to Predict Crime Status," Int. J. Digit. Content Technol. its Appl., vol. 7, no. 9, pp. 361–369, 2013.

[4] S. Amin, "A Step Towards Modeling and Destabilizing Human Trafficking," pp. 2–7.

[5] A. G. Ferguson, "Big data and predictive reasonable suspicision," Univ. PA. Law Rev., vol. 163, no. 2, pp. 329–346, 2012.

[6] U. Saeed, M. Sarim, A. Usmani, A. Mukhtar, A. B. Shaikh, and S. K. Raffat, "Application of Machine learning Algorithms in Crime Classification and Classification Rule Mining," Res. J. Recent Sci. Res.J.Recent Sci, vol. 4, no. 3, pp. 106–114, 2015.

[7] D. Usha and K. Rameshkumar, "A Complete Survey on application of Frequent Pattern Mining and Association Rule Mining on Crime Pattern Mining," Int. J. Adv. Comput. Sci. Technol., vol. 3, no. 4, pp. 264–275, 2014.

[8] S. Schneider, "Predicting Crime: A Review of the Research," 2002.

[9] W. L. Perry, B. McInnes, C. C. Price, S. C. Smith, and J. S. Hollywood, Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. 2013.

[10] L. B. Moses, J. Chan, L. B. Moses, and J. Chan, "Algorithmic prediction in policing : assumptions , evaluation , and accountability accountability," Polic. Soc., vol. 0, no. 0, pp. 1–17, 2016.

[11] M. Camacho-collados and F. Liberatore, "A Decision Support System for predictive police patrolling," Decis. Support Syst., vol. 75, pp. 25–37, 2015.

[12] A. Nasridinov and Y. Park, "A Study on Performance Evaluation of Machine Learning Algorithms for Crime Dataset," Adv. Sci. Technol. Lett. - (Networking Commun. 2014), vol. 66, pp. 90–92, 2014.

[13] C. Yu, W. Ding, P. Chen, and M. Morabito, "Crime Forecasting Using Spatio-temporal Pattern with Ensemble Learning," in Pakdd '14, 2014, pp. 174–185.

[14] E. R. Groff and N. G. La Vigne, "Forecasting the future of predictive crime mapping," Crime Prev. Stud., vol. 13, pp. 29–57, 2002.

[15] W. Gorr, A. Olligschlaeger, and Y. Thompson, "Short-term forecasting of crime," Int. J. Forecast., vol. 19, no. 4, pp. 579–594, 2003.

[16] R. V Clarke and J. E. Eck, "Crime Analysis for Problem Solvers in 60 Small Steps," 2005.

[17] D. J. Paulsen, "Predicting Next Event Locations in a Crime Series using Advanced Spatial Prediction Methods," in UK Crime Mapping Conference, 2005.

[18] V. Grover, R. Adderley, and M. Bramer, "Review of Current Crime Prediction Techniques," in Applications and Innovations in Intelligent Systems XIV, London: Springer London, 2007, pp. 233–237.

[19] M. T. Henderson, J. Wolfers, and E. Zitzewitz, "Predicting Crime," LAW Sch. Univ. CHICAGO, vol. 402, no. April, 2008.

[20] K. H. Vellani, Crime Analysis: for Problem Solving Security Professionals in 25 Small Steps. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc., 2010.

[21] C. Yu, M. W. Ward, M. Morabito, and W. Ding, "Crime Forecasting Using Data Mining Techniques," in 2011 IEEE 11th International Conference on Data Mining Workshops, 2011, pp. 779–786.

[22] A. Galathiya, A. Ganatra, and C. Bhensdadia, "Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning," Int. J. Comput. Sci. Inf. Technol., vol. 3, no. 2, pp. 3427–3431, 2012.

[23] Z. Zhang, "K-means Algorithm." pp. 1–16, 2012.

[24] P. Pitale, A. Ambhaikar, and C. Science, "Prediction tool for Crime Analysis," Int. J. Comput. Technol. Appl., vol. 3, no. June, pp. 1040–1042, 2012.

[25] D. Wang, W. Ding, H. Lo, T. Stepinski, J. Salazar, and M. Morabito, "Crime hotspot mapping using the crime related factors—a spatial data mining approach," Appl. Intell., vol. 39, no. 4, pp. 772–781, Dec. 2013.

[26] A. L. Glenn and A. Raine, "Neurocriminology: implications for the punishment, prediction and prevention of criminal behaviour," Nat. Rev. Neurosci., vol. 15, no. 1, pp. 54–63, 2013.

[27] K. Adhatrao, A. Gaykar, A. Dhawan, R. Jha, and V. Honrao, "Predicting Students' Performance Using Id3 and C4.5 Classification Algorithms," Int. J. Data Min. Knowl. Manag. Process, vol. 3, no. 5, pp. 39–52, 2013.

[28] R. A. Berk and J. Bleich, "Statistical Procedures for Forecasting Criminal Behavior," Criminol. Public Policy, vol. 12, no. 3, pp. 513–544, 2013.

[29] R. A. Berk and J. Bleich, "Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment," Criminol. Public Policy, vol. 12, no. 3, pp. 511–511, 2013.

[30] S. Sathyadevan, D. M. S, and S. G. S., "Crime analysis and prediction using data mining," in 2014 First International Conference on Networks & Soft Computing (ICNSC2014), 2014, pp. 406–412.

[31] D. K. Tayal, A. Jain, S. Arora, S. Agarwal, T. Gupta, and N. Tyagi, "Crime detection and criminal identification in India using data mining techniques," AI Soc., vol. 30, no. 1, pp. 117–127, 2014.

[32] R. Ashok Bolla, "Crime pattern detection using online social media," 2014.

[33] R. J. B. Lehmann, A. M. Goodwill, R. K. Hanson, and K. Dahle, "Crime Scene Behaviors Indicate Risk-Relevant Propensities of Child Molesters," Crim. Justice Behav., vol. 41, no. 8, pp. 1008–1028, 2014.

[34] T. Hart and P. Zandbergen, "Kernel density estimation and hotspot mapping," Polic. An Int. J. Police Strateg. Manag., vol. 37, no. 2, pp. 305–323, 2014.

[35] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland, "Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data," in Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14, 2014, pp. 427–434.

[36] J. Chan and L. B. MOSES, "Using big data for legal and law enforcement decisions : testing the new tools," Univ. N. S. W. Law J., vol. 37, no. 2, pp. 643–678, 2014.

[37] M. Sharma, "Z - CRIME: A data mining tool for the detection of suspicious criminal activities based on decision tree," in 2014 International Conference on Data Mining and Intelligent Computing, ICDMIC 2014, 2014.

[38] R. Kiani, S. Mahdavi, and A. Keshavarzi, "Analysis and Prediction of Crimes by Clustering and Classification," Int. J. Adv. Res. Artif. Intell., vol. 4, no. 8, pp. 11–17, 2015.

[39] F. T. Ngo, R. Govindu, and A. Agarwal, "Assessing the Predictive Utility of Logistic Regression, Classification and Regression Tree, Chi-Squared Automatic Interaction Detection, and Neural Network Models in Predicting Inmate Misconduct," Am. J. Crim. Justice, vol. 40, no. 1, pp. 47–74, Mar. 2015.

[40] R. Pandya and J. Pandya, "C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning," Int. J. Comput. Appl., vol. 117, no. 16, pp. 975–8887, 2015.

[41] R. B. Taylor, J. H. Ratcliffe, and A. Perenzin, "Can We Predict Long-term Community Crime Problems? The Estimation of Ecological Continuity to Model Risk Heterogeneity," J. Res. Crime Delinq., vol. 52, no. 5, pp. 635–657, 2015.

[42] T. Almanie, R. Mirza, and E. Lor, "Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots," Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 4, pp. 01–19, 2015.

[43] W. L. Gorr and Y. J. Lee, "Early Warning System for Temporary Crime Hot Spots," J. Quant. Criminol., vol. 31, no. 1, pp. 25–47, 2015.

[44] P. S. R. Gibbings, D. Chief, S. Richard, T. V. Police, D. S. C. Mcloughlin, and I. A. Ramsey, "Evaluation of the MPS Predictive Policing Trial (June 2015)," 2015.

[45] Z. Hamilton, M. Neuilly, S. Lee, and R. Barnoski, "Isolating modeling effects in offender risk assessment," J. Exp. Criminol., vol. 11, no. 2, pp. 299–318, Jun. 2015.

[46] E. Budur, S. Lee, and V. S. Kong, "Structural Analysis of Criminal Network and Predicting Hidden Links using Machine Learning," CoRR, vol. abs/1507.0, Jul. 2015.

[47] L. Mcclendon and N. Meghanathan, "Using Machine Learning Algorithms To Analyze Crime Data," Mach. Learn. Appl. An Int. J., vol. 2, no. 1, pp. 1–12, 2015.

[48] L. Mcclendon and N. Meghanathan, "Using Machine Learning Algorithms To Analyze Crime Data," Mach. Learn. Appl. An Int. J., vol. 2, no. 1, pp. 1–12, 2015.

[49] A. Gupta, A. Mohammad, A. Syed, and M. N., "A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 7, pp. 374–381, 2016.

[50] V. H. MasÃ-as, M. A. Valle, J. J. Amar, M. Cervantes, G. Brunal, and F. A. Crespo, "Characterising the Personality of the Public Safety Offender and Non-offender using Decision Trees: The Case of Colombia," J. Investig. Psychol. Offender Profiling, vol. 13, no. 3, pp. 198–219, 2016.

[51] N. N. Sakhare, "Classification of Criminal data using J48 Algorithm," IFRSA Int. J. Data Warehous. Min., vol. 4, no. February, pp. 167–171, 2016.

[52] A. K. Singh, N. Prasad, N. Narkhede, and S. Mehta, "Crime: Classification and Pattern Prediction," Int. Adv. Res. J. Sci. Eng. Technol., vol. 3, no. 2, pp. 2393–2395, 2016.

[53] R. R. Brüngger, C. Kadar, I. P. Cvijikj, R. Rosés Brüngger, C. Kadar, and I. Pletikosa, "Design of an Agent-Based Model to Predict Crime (WIP)," in SummerSim-SCSC, 2016, no. June.

[54] M. V. Barnadas, "MACHINE LEARNING APPLIED TO CRIME PREDICTION," 2016.

[55] A. Fine and L. Steinberg, "Self-Control Assessments and Implications for Predicting Adolescent Offending," J. Youth Adolesc., vol. 45, no. 4, pp. 701–712, 2016.

[56] M. A. Tayebi and U. Glässer, Social Network Analysis in Predictive Policing. 2016.

[57] A. S. Akshat Sharma, "Understanding Decision Tree Algorithm by using R Programming Language," in ACEIT Conference Proceeding 2016, 2016, pp. 177–182.

[58] E. Ahishakiye, D. Taremwa, and E. Omulo, "Crime Prediction Using Decision Tree (J48) Classification Algorithm," Int. J. Comput. Inf. Technol. (ISSN2279 – 0764, vol. 6, no. 3, pp. 188–195, 2017.

[59] M. L. Williams, P. Burnap, and L. Sloan, "Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns," Br. J. Criminol., vol. 57, no. 2, pp. 320–340, 2017.

[60] M. S. Vural and M. Gök, "Criminal prediction using Naive Bayes theory," Neural Comput. Appl., vol. 28, no. 9, pp. 2581–2592, Sep. 2017.

[61] P. Olicing, C. R. D. Ata, A. Lgorithms, and E. E. Joh, "FEEDING THE MACHINE," pp. 1–17, 2017.

[62] J. Kleinberg, H. Lakkaraj, J. Leskovec, J. Ludwig, and S. Mullainathan, "Human Decisions and Machine Predictions," NBER Work. Pap. Ser., vol. 23180, p. 76, 2017.

[63] N. Jain, N. Bhanushali, S. Gawade, and G. Jawale, "Physical and Cyber Crime Detection using Digital Forensic Approach: A Complete Digital

Forensic Tool," Int. J. Adv. Res. Ideas Innov. Technol., vol. 3, no. 1, pp. 834–841, 2017.

[64] M. Castelli, R. Sormani, L. Trujillo, and A. Popovič, "Predicting per capita violent crimes in urban areas: an artificial intelligence approach," J. Ambient Intell. Humaniz. Comput., vol. 8, no. 1, pp. 29–36, 2017.

[65] J. Fitterer, T. A. Nelson, and F. Nathoo, "Predictive crime mapping," Police Pract. Res., vol. 16, no. 2, pp. 121–135, 2017.

[66] a Malathi and S. S. Baboo, "Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters," Glob. J. Comput. Sci. Technol., vol. 11, no. 11, 2011.

[67] M. A. Tayebi and U. Glässer, Social Network Analysis in Predictive Policing. Cham: Springer International Publishing, 2016.

[68] N. Andrienko and G. Andrienko, Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach. Springer, 2005.

[69] NIST/SEMATECH, Engineering Statistics Handbook. NIST/SEMATECH, 2012.

[70] W. A. Neil, Introductory Statistics, 5th ed. Addison Wesley, 2002.