How to Cite:
Khalaf, E. F., & Taresh, A. H. (2022). Crime forecasting and prevention using spatial data mining. *International Journal of Health Sciences*, *6*(S8), 2269–2282.
https://doi.org/10.53730/ijhs.v6nS8.12293

Crime forecasting and prevention using spatial data mining

Esraa Faisal Khalaf
Informatics Institute for Postgraduate Studies (IIPS), Iraqi Commission for Computers and Informatics (ICCI), Baghdad, Iraq
Corresponding author email: ms202020598@iips.icci.edu.iq

Dr. Ali Hasan Taresh
University of Information Technology and Communications (UolTC, Baghdad, Iraq
Email: alihtaresh@uoitc.edu.iq

***Abstract*---**This paper focuses on finding safe spot and hotspots using spatial data mining and predicting crime. spatial data mining was applied to a real data set in Los Angeles, and here the study took upon itself the application of the block algorithm using the mean block algorithm K in two stages, the first stage is clustering was performed on the geographical data represented as two features, longitude and latitude, on the coordinates and here represents the spatial data, which is the focus of the study, the result which extract from this stage is , obtain labeled class and to find clusters in the given spatial data in a separate data frame for the coordinators, the second stage, used the K-mode cluster algorithm on the (CRM_DESC, WEAPON_DESC, PREMIS_DESC) classified the crime based on weapon used, premise and crime description to extract classified data under separate crime type data. The last step is to integrate the data extracted from the first stage with the data extracted from the second stage to know the distribution of crimes over Los Angeles regions and counties, in addition to supporting by reviewing the extent of the spread of crimes.

***Keywords*---**spatial data mining, crime forecasting, spatial data.

Introduction

Crime has existed since the dawn of time, and it may be found in any community, regardless of its level of development. People experience terror and untold agony as a result of crime. Crime frequently acts as a roadblock to society's socioeconomic development, discouraging investment, raising transaction costs, and ultimately fueling migration, which leads to global economic inequities.

Banditry, kidnapping, rape, stealing, and murder are only some of the crimes that can be committed. As a result, crime promotes corruption and destabilizes any society's progress. Finally, crime has no boundaries or personalities since it impacts individuals on all levels. If governments do not take strong action to effectively combat the actions of criminals, terrorists, and bandits around the world, this tragic activity will continue to be a major role in the lack of socioeconomic progress [1].

In this paper intend to improve awareness of risky sites in certain time periods by giving a data mining approach using spatial data to determine the most criminal hotspots based on geographical characteristics for a variety of crime categories that is finding hotspots, or places on the map where crime is more concentrated than elsewhere, is the aim of spatiotemporal analysis of crime patterns. Hotspots can exist in a variety of dimensions. They may be zero-dimensional if the crime only takes place in predetermined locations. For instance, individual banks' locations are frequently depicted as dots on a map depicting the locations of bank robberies. A hotspot is a discrete site (like a bank) where crimes frequently occur. In analysis, a hotspot is typically represented on a map as a dot, the size of which is proportional to the number of criminal events there. Furthermore, having this kind of knowledge will assist people in making better living choices. Police forces, on the other hand, can employ this solution to improve crime prediction and prevention. Further-more, this would be beneficial for the allocation of police resources. It can aid in the deployment of police officers to the most likely crime hotspots at any given time, ensuring an effective response. The objectives of this study can be summarized as follows:

- Designing and implementing a prediction system based on learning for forecasting Crime.
- Using spatial data mining and supervised machine-learning algorithm to predict Crime
- A learning algorithm depended on provided input or features that use for train itself into as a functions that could be further used to infer the output or labels.

Problem Statement

Using real-world crime datasets, hope to uncover spatial crime hotspots. Using latitude and longitude geographic data, to locate the most likely crime locations. To predict the type of crime that will happen next in a particular place. Finally, a comparative meta-analytic search was performed based on the findings data with the demographic dataset of those hot or safe areas. The methods mainly identify crime hotspots based on the location of high crime intensity while combating the type of crime.

Table 1

REF	Aim of Study
[2]	Focuses on locating spatial and temporal criminal hotspots. by examining two real-world crime datasets from Denver and Los Angeles. It covers criminal offenses and crime episodes in the city and county of the city for the preceding five calendar years (2010–2015) and provides a statistical analysis supported by many graphs to compare the two datasets.
[3]	Came to know the relationship between Crime rate based on region, monthly levels and data of research was about of last three years of crime (2019 – 2017) and showing crime forecasting trend on data set is extracted from primary data collection based on field work. This dataset consists of about 500 in 10 rows details.
[4]	Another project objective to develop a ML model for crime predicting based on geographical characteristics for a variety of crime categories. (OSM) spatial data is returned using the reverse geocoding approach. This research is also aimed at identifying hotspots. Then, based on the location of distinct hot spots for various forms of crime, the spatial distance feature was generated, and this value was employed as a feature for classifiers.
[5]	In order to anticipate future crimes, suggest a machine learning model that learns from prior crimes. Based on past crime data, the goal of this work is to forecast criminal activity for a certain day of the year and at a particular location in the city. In other words, it entails choosing a model that is determined by a certain city location and time of day. On crime dataset from the city of Chicago in the United States of America that were able to retrieve through the city's website mentioned in. This data set contains 6755055 lines or crimes of 18 years from 2001 to 2018 and with 22 columns of information.

Dataset

In this study, used datasets from real-word crimes in Los Angeles. To construct spatial data mining models, mainly focused on dataset of crime the period 2020 to Present. The crime data provided by Los Angeles Police Department includes reported cases this dataset reflects incidents of crime in the City of Los Angeles. see Fig. 1. This dataset is composed of 28 attributes with 458547 instances [8]. The key attributes provide, the following shows the used key attributes and its content values Table II.

Table II

Attributes	Data Type	No. of Distinct Value	Values
Crm Cd Desc	object	134 Categories	VEHICLE – STOLEN BATTERY - SIMPLE ASSAULT BURGLARY FROM VEHICLE THEFT FROM PERSON – PETTY THEFT … etc
WEAPON_DESC	object	78 Categories	SCREWDRIVER HAMMER

			FIXED OBJECT TOY GUN STRONG-ARM (HANDS, FIST, FEET) ... etc
PREMIS_DESC	object	306 Categories	SINGLE FAMILY DWELLING STREET PARKING LOT SIDEWALK ... etc
LAT	float64	Limited (34.0141, 34.3226)	
LON	float64	Limited (-118.297, -118.5426)	

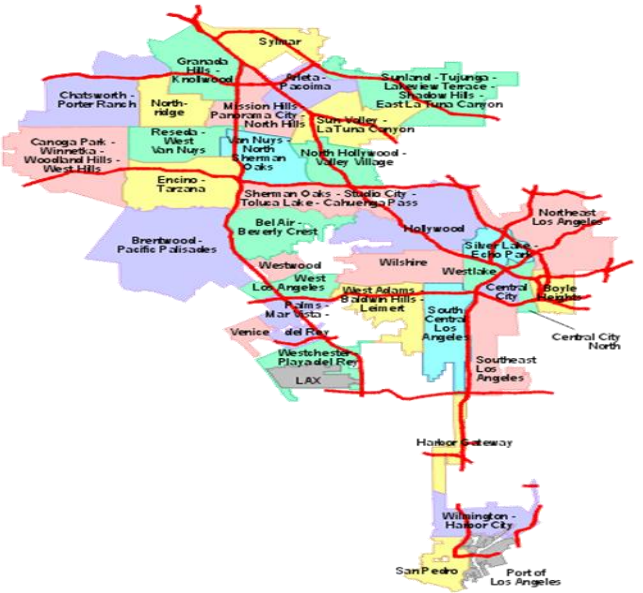

Fig. 1. Map of Los Angeles areas [6]

Methodology

Through this search, conducted an analysis of a dataset of crimes from the real world of Los Angeles, here the search took upon itself the application of one algorithm of data mining cluster using spatial data in two stages, the first phase performed clustering on geographic data represented based on two features, longitude and latitude, (the coordinates) and here the spatial data represents , which is the focus of the search, which extract from this stage, to extract labeled data in a separate data frame for the coordinates, the second stage, used the same cluster algorithm, but here on the crime type attribute to extract labeled data in a separate data frame for crime type. The last step is to combine the data extracted from the first stage with the data extracted from the second stage to see the distribution of crimes on the regions and counties of Los Angeles, in addition, support by with predictions model extent. see Fig. 2 describe this abstract scenario.

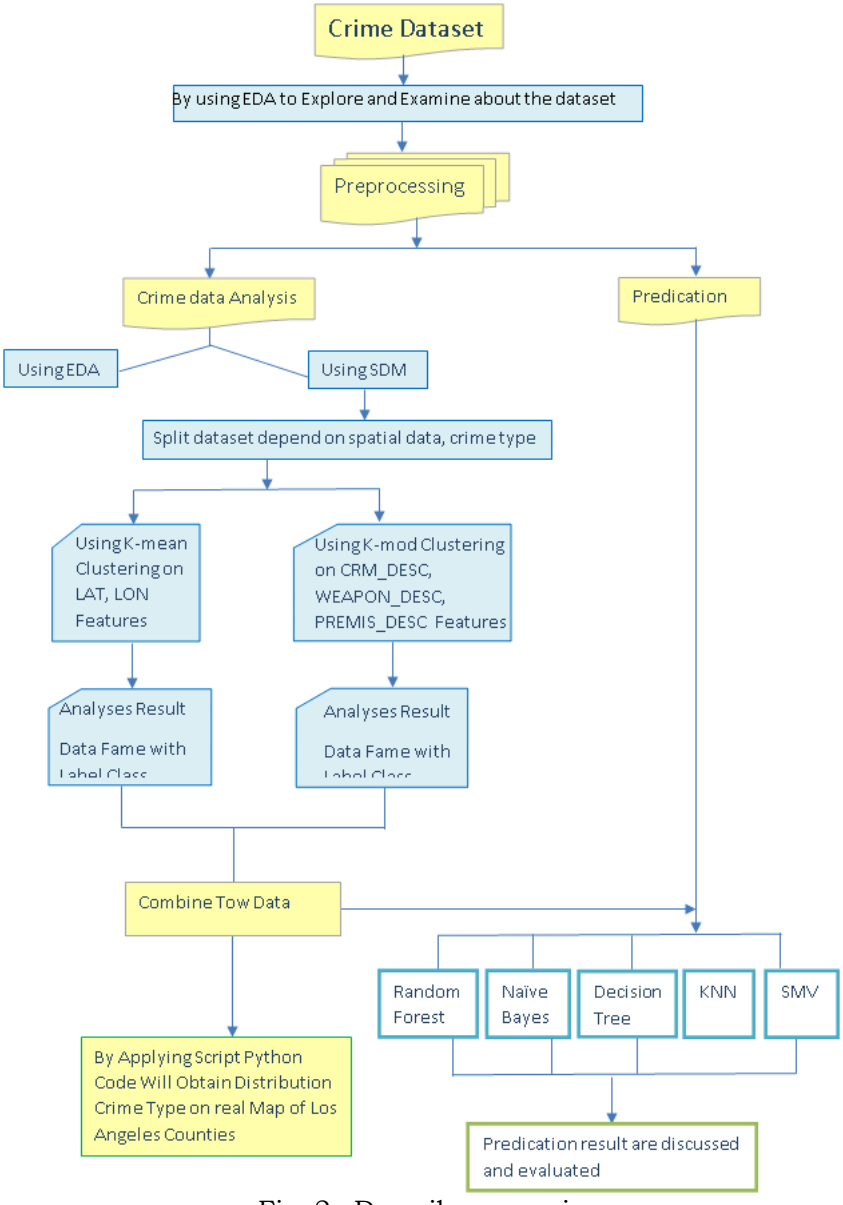

Fig. 2. Describe scenario

Data Preprocessing

Preprocessing steps were applied to prepare the dataset in an appropriate form. performed the following preprocessing steps on the selected datasets:

Dropping Unnecessary Columns

Often, you'll find that not all the categories of data in a dataset are useful to perform job. In this case, retaining these unneeded categories will take up unnecessary space and potentially also bog down runtime. Some columns will not

assist us in data mining analyses, thus decided was to drop them. In detail the DR_NO is a unique code, area code, district code, date reported, crime code, part1/2, premises code, incident code, upon used code and status code. Pandas library provides a handy way of removing unwanted columns or rows from a data frame with the drop() function

Data Cleaning

Using .str() methods to split Date OCC Column to Year, Month, Day each of them separated column. see Fig.3

Out[397]:

	DATE OCC	Year	Month	Day
57910	2020-07-12	2020	July	Sunday
6076	2020-09-07	2020	September	Monday
288174	2021-06-01	2021	June	Tuesday
15131	2020-04-25	2020	April	Saturday
175009	2020-12-28	2020	December	Monday

Fig. 3. Split DATE OCC column to Year, Month, Day

Handling Missing Value

Most of the machine learning models as well as the data mining algorithms seek to use will lead us to error results if pass them with NaN values. The easiest way that comes to mind is to try to replace it or populate it with 0, but this can greatly reduce the accuracy of your model. After used script code python to visualize the data frame see there are missing value in some column like the figure (3.2). Pandas library provides a handy way of removing unwanted columns data frame with the drop() function, but if rows has missing value handling by replaced each NaN value with Unknown if the data type of column is object , and replace NaN value with 0 if the data type of column is numeric (float or integer) so not drop any row from a data frame with the drop() function. But when dealing with data set through applying data mining algorithm or machine learning algorithm will exclusion each value is equal to unknown or 0. using EDA to visualize which column has missing value Fig. 4. Then Fig. 5, result of handling missing value.

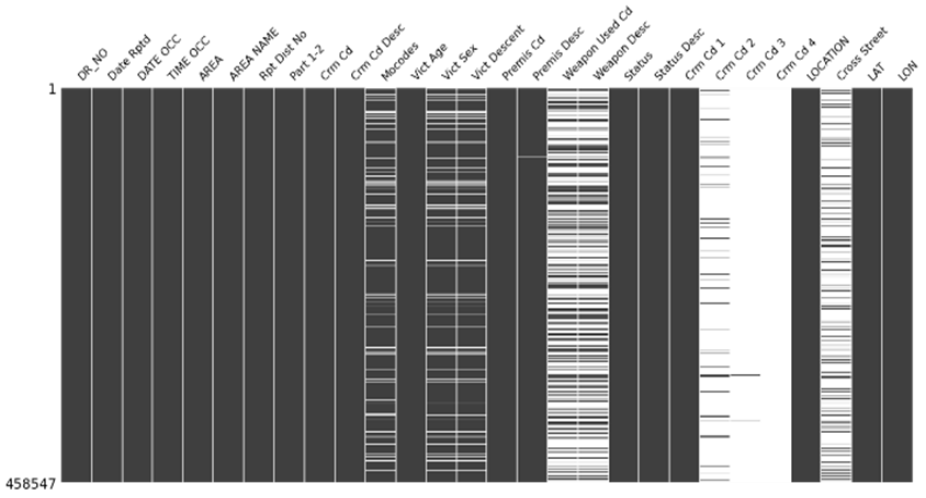

Fig. 4. Visuale missing values in Los Angeles dataset

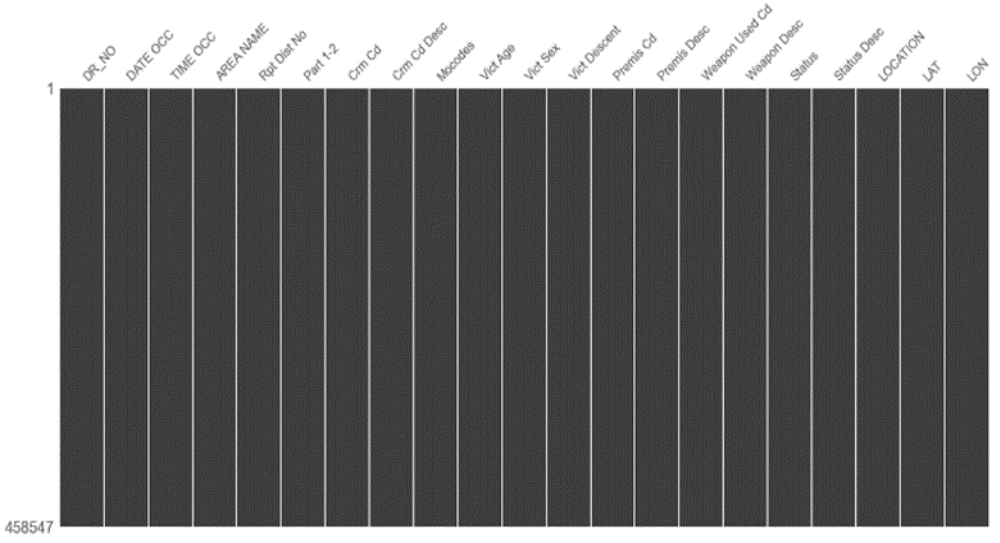

Fig. 5. Visuale handling missing values in Los Angeles dataset

Data Analysis

Data Analysis Using Exploratory Data Analysis (EDA)

Conducted an initial analysis using Exploratory Data Analysis (EDA) of data as a first step. see how the crime distribution on Los Angeles map by using key features (Crm Cd Desc, Year, LAT, LON). Fig. 6.

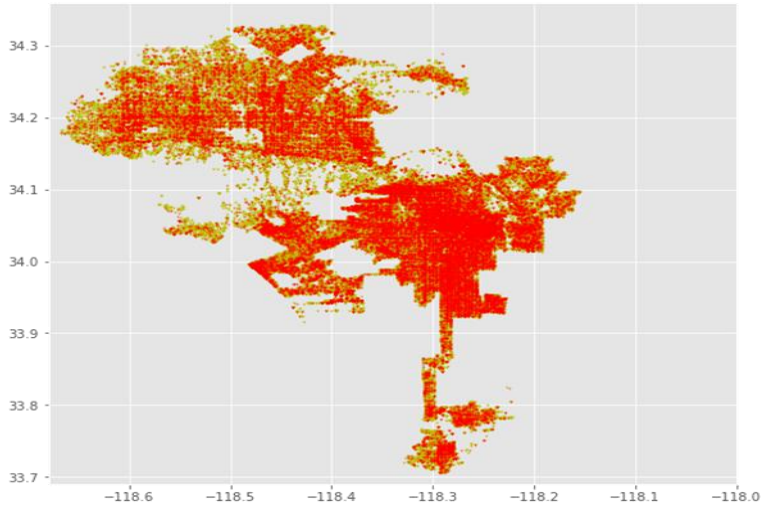

Fig. 6. Crime distribution on Los Angeles map

Data Analysis Using Spatial Data Mining (SDM)

After the completion of the data processing process mentioned above and the progress of the analysis process using the results were reviewed, now came the phase of analysis using spatial data mining (GEOGRAPHICAL CLUSTERING). The first step is get the latitudes and longitudes as separate data frame, then plot the data of longitude and latitude, clearly see that the data has taken the shape of LA, Fig. 7

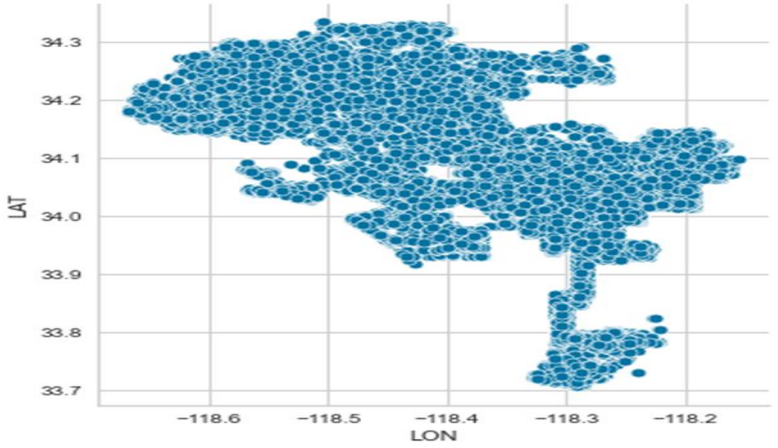

Fig. 7. Data of longitude and latitude taken the shape of LA Map

Transform the data so that all features have equal variance. This is a necessary step for all clustering analysis. Use the elbow method to find the best number of clusters for our k-mean clustering, see that the optimal number of clusters to use for this analysis is 5. That means the best number of groups in order to create groups with members that are more similar between them than with members of other groups is 5, Fig. 8.

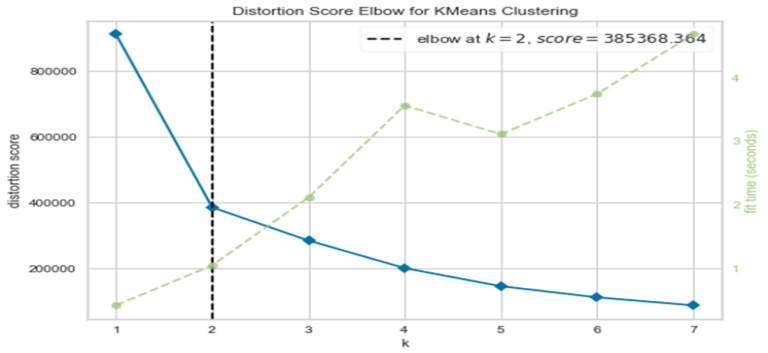

Fig. 8. Elbow method Graph

Execute the k-means clustering with 5 centroids. Fig. 9.

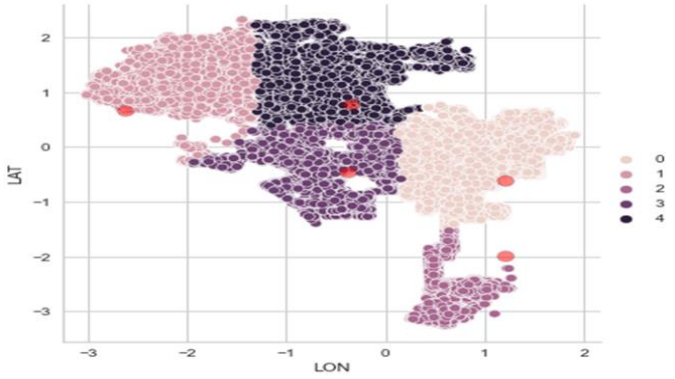

Fig. 9. Execute the k-means clustering with 5 centroids

Transforming back the centroids in order to show them on the map. Now can see the results on the LA map, (Fig.10).

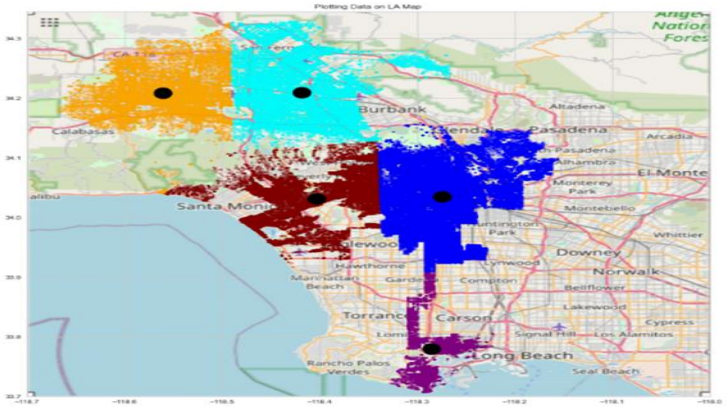

Fig. 10. Clustering Los Angeles counties map

Now come to the result of the second stage of the phase analysis using data mining, which is the classification of crimes away from the traditional

classification by the type of crime, the method which used to classify the crime is according to the type of weapon used, the crime scene and the description of the crime cluster by using categorical variables use k-modes to do that on features (CRM_DESC, WEAPON_DESC, PREMIS_DESC), at the beginning separated each of them as data frame, try clean them from attributes that rarely appear, then merge all the above data frame into a new data frame which contains the most occurred crime types, weapons used and premises, and execute the k-modes algorithm with 10 different clusters (groups). Then Create a new data frame with the district codes.

Now use cluster analysis to on data frame with the district codes to find groups with districts that are most similar to each other and most different to districts of other clusters when it comes to the types of crimes committed , use StandardScaler() to transform the data so that all columns have equal variance. Necessary step in clustering. In detail, the mean of the distribution is 0 and the variance is equal to 1. shown in Fig.11

	DISTR_NO	label
0	101	2
1	105	2
2	109	2
3	111	0
4	112	2
...	...	...
1179	2189	1
1180	2196	2
1181	2197	2
1182	2198	2
1183	2199	2

1184 rows × 2 columns

Fig. 11. DISTR_NO with labeled

Finally combine the result of cluster of first stage with the result of first stage to obtain data frame that includes labels, district and longitude and latitude for Fig.12

	DISTR_NO	label	LAT	LON
0	101	1	34.0677	-118.2398
1	101	1	34.0652	-118.2451
2	101	1	34.0652	-118.2468
3	101	1	34.0685	-118.2460
4	101	1	34.0674	-118.2468
...	...	...	...	...
458542	2198	1	34.1480	-118.6019
458543	2198	1	34.1480	-118.6019
458544	2198	1	34.1482	-118.6052
458545	2198	1	34.1480	-118.6019
458546	2199	1	34.1356	-118.5710

Fig. 12. Combine the result of cluster of first stage with the result of stage

Can see result use cluster analysis to find groups with districts that are most similar to each other and most different to districts of other clusters when it

comes to the types of crimes committed, will use StandardScaler() to transform the data so that all columns have equal variance. Necessary step in clustering. In detail, the mean of the distribution is 0 and the variance is equal to 1.

After use k-means clustering to grouping the data, by checking the centroids centers, can understand the type of district it describes.

- In details, the first centroid has values below 0 for all the columns which means this centroid describes the safer district, meaning the ones with the lesser crime types reported.
- The second centroid has values above 1 in most columns which means this centroid describes the more dangerous district, meaning the ones with the most crime types reported.
- Lastly, the third centroid has values slightly above (sometimes below) 0 for all the columns which means this centroid describes the medium safe district, meaning the ones with the average crime types reported. The result can see on the map, shown in Fig.13.

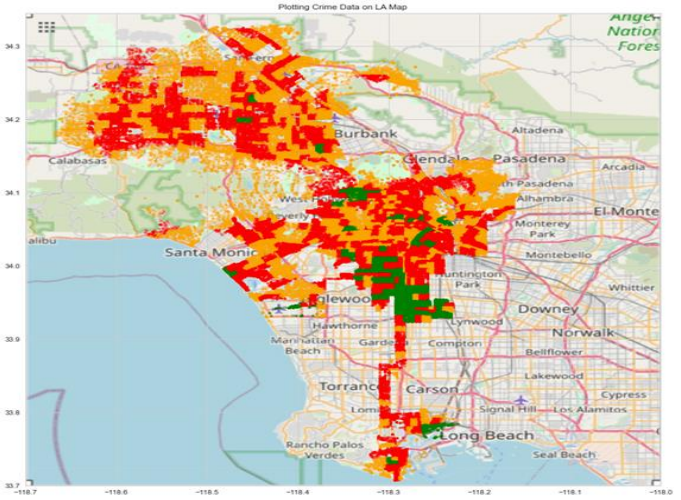

Fig. 13. Showcase the results of clustering on map

In the map can see three different colors of points. Every point is a reported crime incident from 2020 till 2022. Firstly, grouped by each district taking the sum of counts for each different crime type committed in that district. Then clustered the districts into three different clusters. After finding out that each different cluster defines the safety of the district because of the values of the centroids can make conclusions on the safety of the regions:

- The green points of the map concern districts with lesser crimes that normally. Those are the safer zones. That is because those districts belong to the clusters with the negative values of the centroid. Some examples are: Bel Air, Beverly Hills, West Hollywood, Brentwood, Mulholland / Sepulveda Blvd and Palisades Ave De Santa Ynez
- The orange points of the map concern districts with average amount of crimes. Those are neither the safest nor the most dangerous zones. That is

explained from the fact that those districts belong to the clusters with the values of the centroid that are approximately 0. Some examples are: Harbor Blvd, Forest Lawn Dolanco Junction and Tampa Ave

- The red points of the maps concern districts with more crimes that normally, those are considered the dangerous zones. That is because those districts belong to the clusters with the high (positive) values of the centroid. Examples are Downtown, South Park, Central city and Fashion District

The regions that have no color are either very safe (no crimes committed) or there are no data crimes reported or the crimes incidents are missing. For example, for the Marvin Braude Mulholland Gateway Park (located on the north west LA, green forest) can hypothesize that no crimes were committed there because people are not living there. However, for the area of Torrance can hypothesize that the crimes for that region are missing because there are no incidents.Finally, can see the crime areas are highlighted on Los Angeles map with count of crime for each area, shown in Fig.14

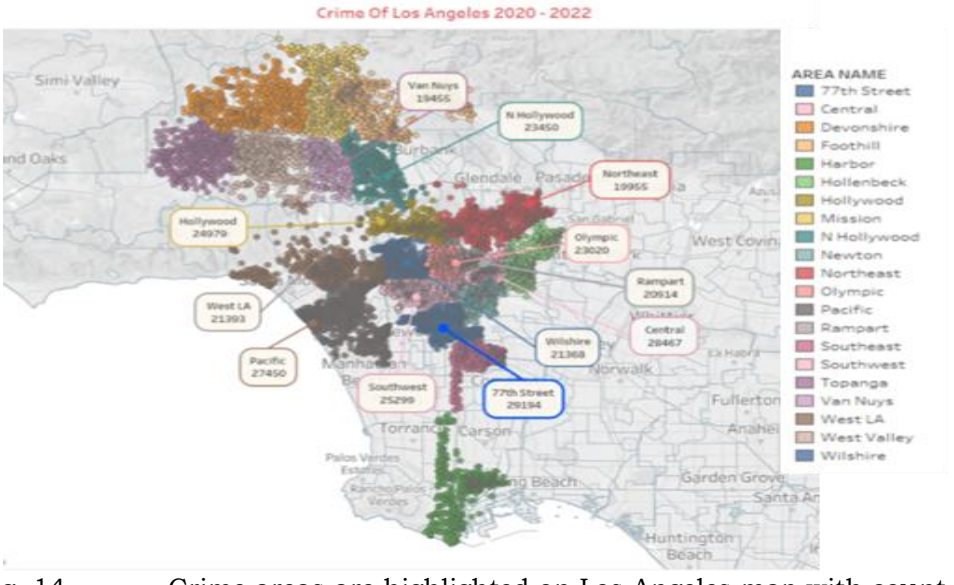

Fig. 14. Crime areas are highlighted on Los Angeles map with count of crime for each area

Predicate Model

examples or training units are given the proper outputs, and the algorithm learns to respond more precisely by comparing its outputs to those provided as inputs on the basis of these training sets. Learning under supervision is often referred to as learning through models or by using examples.

Result of Predicate Model

performance criteria like accuracy, Training time foreach model see Table III.

Table III Result of Predicate Model

Model	Accuracy w/o scaling	Training time (sec)
Random Forest	1.000000	5.31
Naïve Bayes	1.000000	0.03
Decision Tree	1.000000	0.72
KNN	0.998175	0.26
SVM	0.518032	46.36

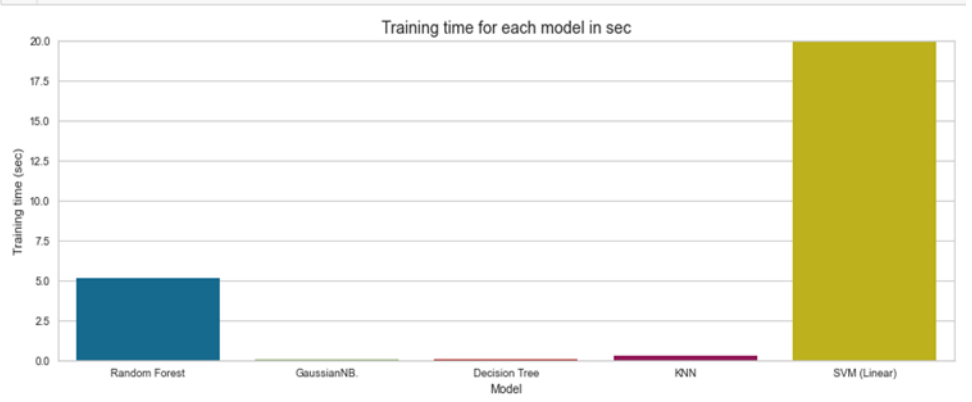

Fig. 15. Accuracy of the result of predicate model

Fig. 16.

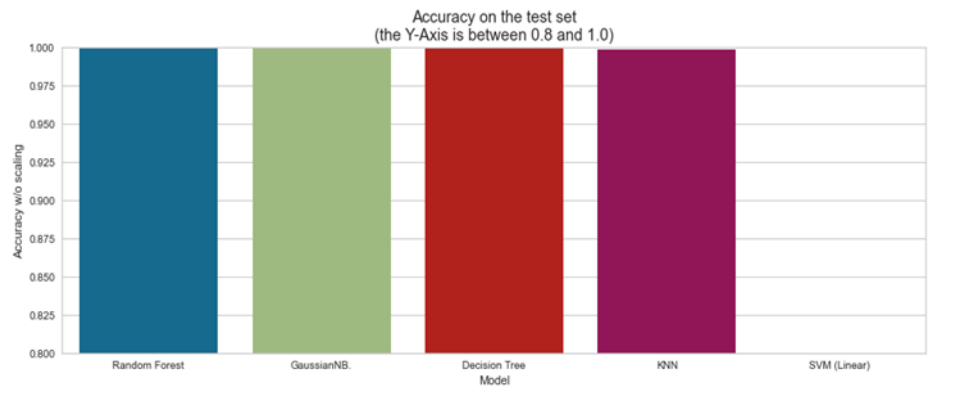

Fig. 17. Training Time of the result of predicate model

Conclusions

In terms of application, the study of the analysis relates the data from this algorithm to the actual sites of crime activities throughout the target city. This shows that the proposed use of geographical data has a lot of promise in terms of anticipating crime and how it is dispersed on a map. Our research's ultimate goal is to deliver prediction in the proper place with minimal data input. In the end, law enforcement is able to tackle criminals proactively rather than reactively.

References

1. A. Soares, “Predicting Crime Using Spatial Features,” no. March, 2018.
2. Alghamdi, A. G., AlZain, M. A., & Masud, M. (2022). Cloud computing environments and management for big data security, and performance (CBDS) Model. *International Journal of Health Sciences*, *6*(S1), 8860–8878. https://doi.org/10.53730/ijhs.v6nS1.7030
3. Available:https://statisticalatlas.com/United-States/Overview.
4. H. Aitelbour, S. Ounacer, Y. Elghomari, H. Jihal, and M. Azzouazi, “A crime prediction model based on spatial and temporal data,” no. November 2018, 2019, doi: 10.21533/pen. v6i2.524.
5. https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8.
6. Laalmanac.com, 'City of Los Angeles Planning Areas Map', 2022. [Online]. Available: http://www.laalmanac.com/LA/la00a.php.
7. S. Mahmud, M. Nuha, and A. Sattar, Crime Rate Prediction Using Machine Learning and Data Mining, vol. 1248, no. June. Springer Singapore, 2021.
8. Simpen, I. N., Redana, I. W., & Umratul, I. (2018). Aquifers selection based on geoelectric method data in the framework of drilling wells: A case study on international hospital project in Nyitdah Tabanan Bali. *International Journal of Physical Sciences and Engineering*, *2*(2), 68–78. https://doi.org/10.29332/ijpse.v2n2.151
9. T. Almanie, R. Mirza, and E. Lor, “c,” Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 4, pp. 01–19, 2015, doi: 10.5121/ijdkp.2015.5401.
10. U. Thongsatapornwatana, “A survey of data mining techniques for analyzing crime patterns,” 2016 2nd Asian Conf. Def. Technol. ACDT 2016, pp. 123–128, 2016, doi: 10.1109/ACDT.2016.7437655.
11. Widana, I.K., Sumetri, N.W., Sutapa, I.K., Suryasa, W. (2021). Anthropometric measures for better cardiovascular and musculoskeletal health. *Computer Applications in Engineering Education*, *29*(3), 550–561. https://doi.org/10.1002/cae.22202