

PROFESSIONAL TRAINING REPORT
at
Sathyabama Institute of Science and Technology
(Deemed to be University)

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Technology Degree in

Information Technology

By

DEWASHISH KUMAR (37120021)

ABHINASH KUMAR (37120002)



DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF COMPUTING
SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600119. TAMIL NADU

MARCH 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with "A" Grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600 119.

Phone: 044 - 2450 3150 / 51 / 52 / 54 / 55 Fax: 044 - 2450 2344

www.sathyabama.ac.in



SCHOOL OF COMPUTING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **DEWASHISH KUMAR (37120021) & ABHINASH KUMAR (37120002)** who carried out the “**PREDICTING THE RATE OF CRIME USING KNN ALGORITHM**” under our supervision from November 2020 to April 2021.

Internal Guide

Dr. S.Gowri, M.E., Ph.D.,

Head of the Department

Dr. R.SUBHASHINI, M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I, DEWASHISH KUMAR (37120021) and ABHINASH KUMAR (37120002) hereby declare that the Professional Training Report on “**PREDICTING THE RATE OF CRIME USING KNN ALGORITHM**” done under the guidance of **Dr. S.GOWRI, M.E., Ph.D.**, at Sathyabama Institute Of Science And Technology has submitted in partial fulfillment of the requirements for the award of Bachelor of Technology degree in Information Technology.

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind Encouragement in doing this project and for completing it successfully. I am grateful to them

I convey my thanks to **Dr. T.Sasikala M.E., Ph.D, Dean**, School of Computing , **Dr.SUBHASHINI M.E., Ph.D.**, Head of the Department of Information and Technology for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. S.Gowri, M.E., Ph.D.**, for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Information Technology** who were helpful in many ways for the completion of the project.

ABSTRACT

Crime is one of the greatest and ruling issue in our general public and its anticipation is a significant task. Every day there are gigantic quantities of violations submitted much of the time. This require monitoring all the violations and keeping up a database for same which might be utilized for future reference. The present issue confronted are keeping up of legitimate dataset of wrongdoing and dissecting this information to help in foreseeing and fathoming violations in future. The goal of this task is to break down dataset which comprise of various violations and foreseeing the kind of wrongdoing which may occur in future relying on different conditions. Right now, will utilize the strategy of AI and information science. Before preparing of the model information preprocessing will be finished after this component determination and scaling will be done with the goal that precision get will be high. The K-Nearest Neighbor (KNN) grouping and different calculations will be tried for wrongdoing expectation and one with better exactness will be utilized for preparing. Perception of dataset will be done as far as graphical portrayal of numerous cases for instance at which time the criminal rates are high or at which month the crimes are high. The spirit reason for this task is to give a quip thought of how AI can be utilized by the law implementation organizations to recognize, foresee and illuminate violations at an a lot quicker rate and in this manner decreases the crime percentage.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	v
	LIST OF FIGURES	vii
	LIST OF ABBREVIATIONS	vii
1	INTRODUCTION	1
	1.1 ABOUT THE PROJECT	1
2	LITERATURE SURVEY	2
	2.1 RELATED WORK	2
	2.2 SURVEY ON CRIME	2
3	METHODOLOGY	4
	3.1 PROPOSED MODELLING	4
	3.2 CLUSTERING ALGORITHM	6
	3.3 STREAMING ALGORITHM	8
	3.4 SYSTEM CLASSIFICATION	9
	3.5 SYSTEM REQUIREMENTS	10
4	RESULTS AND DISCUSSION	12
	4.1 RESULTS	12
	4.2 DISCUSSION	12
5	CONCLUSION AND FUTURE WORK	13
	REFERENCES	14

APPENDICES	15
A. SAMPLE CODE	15
B. SCREEN SHOTS	18

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO.
2.1	TABULATION OF SURVEY ON CRIMES	3
2.2	BAR GRAPH REPRESENTATION OF CRIMES	3
3.1	ARCHITECTURE DIAGRAM	5
3.2	GRAPH OF K-MEANS CLUSTERING	7

LIST OF ABBREVIATIONS

- 1 SVM – Support Vector Machine
- 2 NB – Naive Bayes
- 3 Py – Python
- 4 Sklearn – Scikit learn
- 5 Np – Numpy

Chapter-1

INTRODUCTION

1.1 ABOUT THE PROJECT

Crimes are normally based on a public drawback that affects the society and therefore the profit-making process of the culture. There is plenty of amendment within the crime pattern and this makes enforcement agencies to use data processing nearby to the enhance crime analytics and for higher growth of groups in form of communities. Hotspot analysis guides us to spot the foremost seemingly criminal hotspots in a very specific region, and detects that location have intensity of crime as more. By victimization this technique, individuals will alert themselves concerning the damaging location before reaching a specific destination. This helps individuals to be conscious of the place wherever the incidence rate of illegal activities are outstanding. This type of knowledge concerning the whereabouts of the crime and improves the living selections of individuals. Police uses these leads to crime prediction and to prevent the crime. Personnel are often utilized at the most seemingly crime places for the betterment of individuals. The knowledge out there when attending crime analysis will build the individuals to be safe.

Computer based intelligence is a subfield of programming designing, anyway is every now and again moreover insinuated as perceptive examination, or insightful illustrating. Its target and use is to create new or possibly impact existing counts to pick up from data, in order to fabricate generalized models that give careful desires, or to find plans, particularly with new and disguised tantamount data. Toward the beginning of an AI adventure, a dataset is typically part into a couple of subsets. The base subsets are the planning and test datasets, and consistently an optional third endorsement dataset is made moreover. At the point when these data subsets are produced using the basic dataset, a judicious model or classifier is readied using the planning data, and a short time later the model's insightful exactness is settled using the test data.

Chapter-2

LITERATURE SURVEY

2.1 RELATED WORK

The application of information preparing on criminal systems contains a fierce potential for the information driven battle against the composed and productive crime, hole new difficulties inside the field of interpersonal organization investigation. The formulation of crime pattern detection uses the machine learning techniques and thereby use data processing to support police officers in finding crimes. The options for Crime Hotspot Detection and Crime class Classification mistreatment Machine Learning techniques tested based on many years of crime records. The bike and car robbery hotspots were learned in business space parking, hospital parking for emergency, and on roadsides , Snatching hotspots were found inside the thickly possessed neighborhoods lacking road lights, and furthermore the hotspots of thefts were discovered inside the area of town. To guarantee sensible use, they adopted an internet service design as the simplest way to integrate with public safety systems and platforms, as ROTA one. Through knowledge clump, individuals will realize the data distribution, character observes of all clusters, and creates additional study of the particular clusters. The point in time clump shows that polydimensional to outlaw profession successively to supply a representation clustering report of categories of those shocking activities. Exemplar-based clump strategies shows the way to manufacture progressive results on variety of artificial and real-world clustering issues. they are appealing as a result of they provide process edges and might handle absolute pairwise similarity measures between knowledge points.

2.2 SURVEY ON CRIME

The public surveys take the full report on the number or amount of crimes that has not reported to the police .This surveys are usually more useful for assessing data. Statistics for various types of crime in the country are recorded and kept for reference for the officials. The survey of crimes in 2016,2017 and 2018 has been recorded .Crime types are analyzed and the number of crimes encountered are reported and the total

number of crimes are also noted.

1	CRIME TYPE	CRIME 2016	CRIME 2017	CRIME 2018	TOTAL
2	BIKE THEFT	679	567	398	1644
3	CAR THEFT	201	102	92	395
4	ROBBERY	231	307	395	933
5	SNATCHING	244	105	126	475
6	RAPE	378	456	571	1405

Table 2.1:Tabulation of survey on crimes

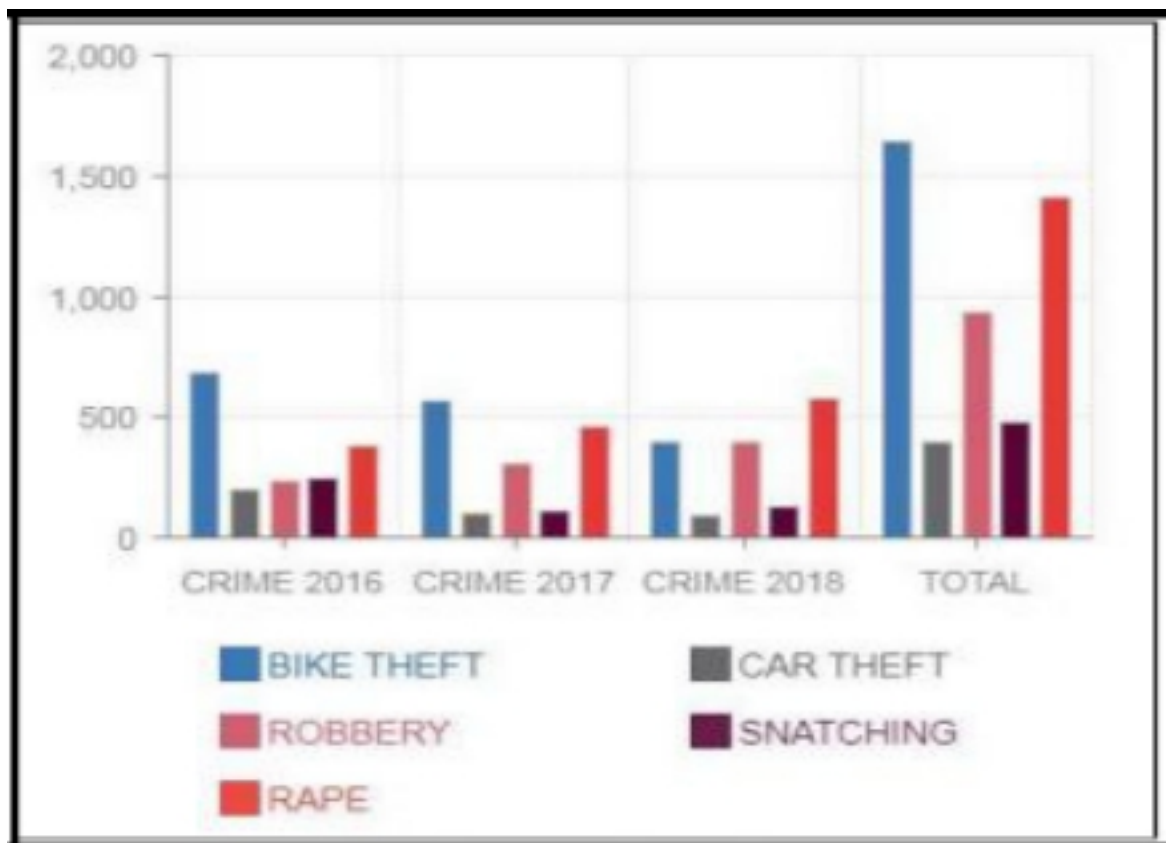


Figure 2.2:Bar Graph Representation of crimes

Chapter-3

METHODOLOGY

3.1 PROPOSED MODELLING

Crime knowledge analyst wants smart domain learning and mapping real data to data processing attributes isn't invariably a simple task and sometimes needs masterly data miner this is often matched by data stream mining. Process of gathering data structures from continuous, speedy data records is named data stream mining. An information stream is an ordered sequence of example within which several applications of information stream mining will be scan one time or some variety of times victimisation restricted quantity of computing and storage capability. This paper proposes and explains an ensemble-based stream mining algorithmic rule supported supervised learning that tells the challenge by maintaining an evolving assortment of multiple models to classify energetic data streams of infinite length. During this paper we worked on three parameters for detection and identifying criminals in the criminal network. Moreover, an algorithmic rule is developed and therefore the adopted metrics are well-known in an exceedingly streaming data atmosphere. The tested analysis are distinguished and the ancient and new parameters are obtained in predicting and detecting criminal activities that replies on our DSP Application .

A. Crime branch Database: An collection of information that is composed with the goal that it very well may be effectively accessed, managed and updated which is called crime branch database. New information can be updated, managed and deleted from the database . This includes all the information about the crimes that happen in the area or region.

B. Factors Extraction: The main factors required for clustering the data are gathered.

C. Criminal Profile: Crime suspected have there profile which includes all the details of the person related to their background.

D. Analysis: All the values from the criminal profile data and from the factors extracted are analyzed.

E. Profile Creation: Now, the profile is created about the crime suspect and the user can access to that account.

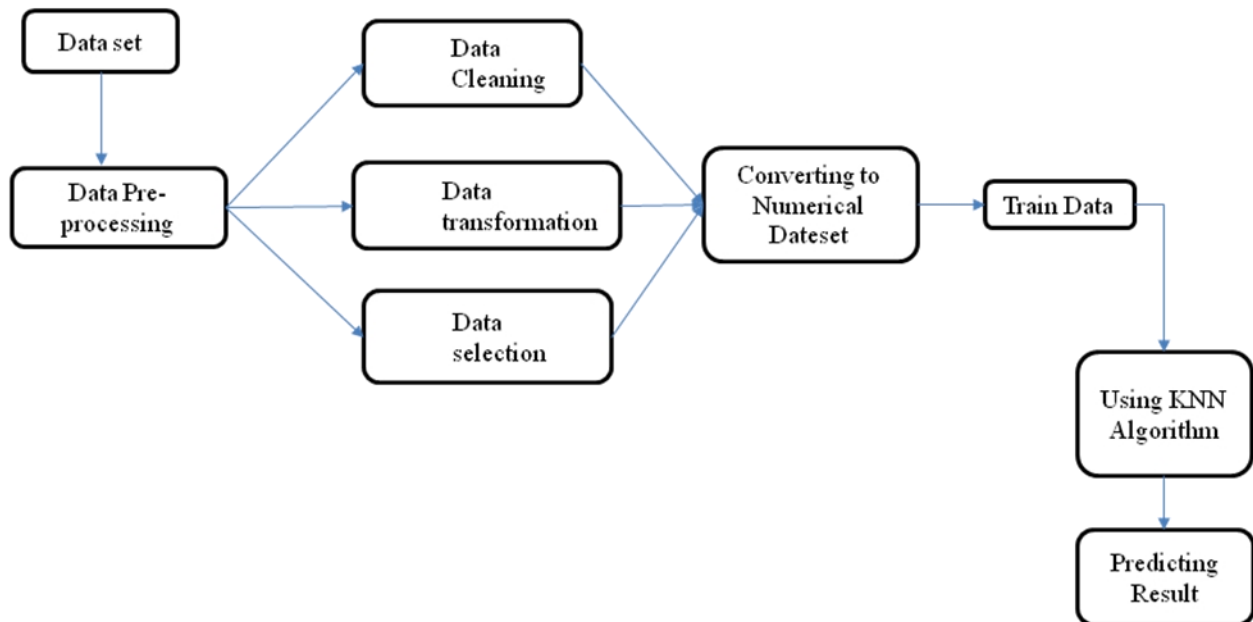


Figure. 3.1. Architecture Diagram

F.Profile Activities: This gives the details of what crime the suspect did. All the crime activities are shown in profile activities.

G.Profile Difference: The obtained profile difference for the various styles of crimes supported records from the police, they are classified as traffic, bike theft and robbery, kidnap and different sexual crimes. Every crime profile has some intrinsic seriousness pattern. This explains the amount of crimes and notice number of proportion of the crimes happening once a year that is classified and handling illegal activities seriousness by aggregation them into criminal profiles.

H.Clustering: This involves the grouping of data points, which is a machine learning technique. To classify each data point into a specific group by using crime data we can also use clustering algorithm.

I.Streaming Algorithm: It is the method of extracting crime data from continuous, rapid data records or informations. All the data are streamed and the user can view the live data about the crime that happens in that area or locality.

3.2 CLUSTERING ALGORITHM

Clustering algorithm has become a rising analysis issue in data processing field because of its kind of applications. several knowledge cluster algorithmic program during this recent years has advent many data and its value additional employed in vast kind of application as well as image process, process micro-objects, visual converse, drugs or political economy, leads the quality of the algorithms. The major downside behind the knowledge cluster algorithms is that it cannot be in order. Algorithmic program progress might offer glorious results as a base with knowledge set even though this won't give proper results. Clustering algorithms are mostly learning which are unsupervised. It is almost preprocessing of data. They are actually updated on day to day basis. Only some can actually work for some data. Clustering in data mining is a very useful algorithm when it comes to cluster many data items and even the result can be obtained in scalar form.\

Steps to follow for K-means algorithm:

Let us assume X = be the vector for data clusters and V = be the points of centres represented as sets..

- 1) Clustering centres are considered as 'c'.
- 2) Compute the gap within the cluster centers and every data points.
- 3) Mark the information purpose to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

$$v_i = (1 / c_i) \sum_{j=1}^{c_i} x_i$$

- 4) compute using the new cluster center using:
where, 'ci' represents the amount of data in ith cluster.
- 5) compute the gap within the cluster centers and every data points.
- 6) If no data point was reassigned then end or otherwise repeat from step 3.

K-means is one of the best way of solving the clustering problems. The centroids of k clustering which can label new data entries. This method is to classify a given data set

through a particular technique of clusters (assume k clusters) mounted. The foremost set up is to stipulate k centers, every for one cluster. The centers should be placed during a correct manner because of absolute distinction in location and totally different result. the higher alternative is to position them as greatly like doable distant from each other clusters. the following step is to want each purpose happiness to a selected information set and associate it to the nearest center. once no purpose is unfinished, the next step is approached. At presently we prefer to re-calculate k new centroids. A loop is generated. Thus, the k centers change the locations one by one until no several of the changes are done. Thus, this formula aims on minimizing associate objective operate grasp as square error operate that's given by:

$$J(V) = \sum_{i=1}^c \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

where, $(\|x_i - v_j\|)$ is the Euclidean distance between x_i and v_j .
 'c' is the number of cluster centers.

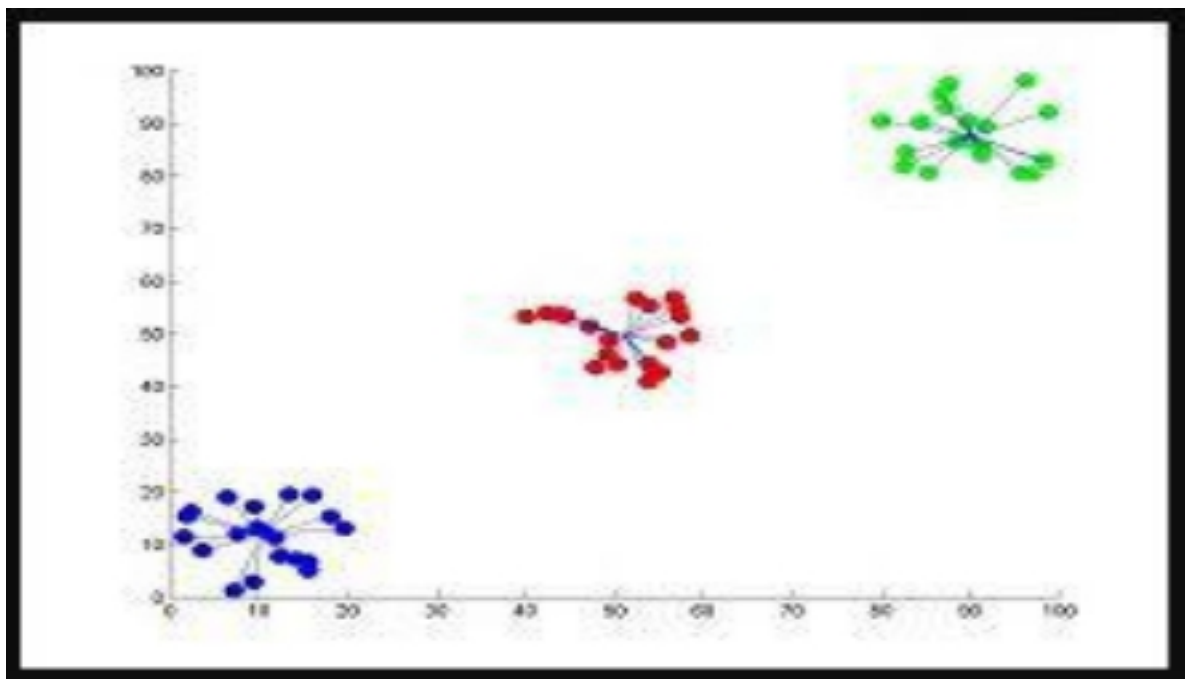


Figure 3.2 : Graph of K-Means Clustering

3.3. STREAMING ALGORITHM

The streaming algorithm is used for the large amount of data that are added in live and people can access the data and can predict which area is most dangerous in the case of illegal activities. By the data from enforcement agencies, personnel can know about the whereabouts of the criminals and by the concept of streaming data mining one can easily cluster the data and the patterns for the data are analyzed and the link between each criminal activity is also monitored. About the crime suspects and therefore the ability to find patterns in an exceedingly (near) period manner will be of primary importance. The patterns for link detection and prediction on an evolving criminal network are known. For capturing live data and processing it we need to work on with live data streaming technologies. Criminal based records for the particular region can be analysed without any delay.

Data Stream Mining is that the gathered structures from continuous, fast data informations in form of records. Data stream is associated ordered sequence that are used in many applications of data stream mining is also browse simply the once or atiny low style of times practice restricted computing and storage capabilities. Locations in map with more amount of crime intensity are mentioned crime hotspots. Even before definition of hotspot has not been found however, it will simplistically be explained as part that contains a heap of occurrences of criminal incidents as compared to its surrounding areas. Scientist's investigation the utilization of these hotspots specifically zones and also the explanation for its commonness, when that, the investigators turn out and utilize distinctive strategies to examination the data set thereby created. Individuals, notably examiners and analysts United Nations organization share an energy for news identified with sociology can demonstrate pride in wrongdoing hypotheses, because of it's been noticed that they assume a noteworthy job in clarifying the explanation for the augmentation of rate in explicit zones anyway in light of the fact that the decrements of rate in certain areas.

3.4. SYSTEM CLASSIFICATION

Albeit regulated and unaided learning are two of the most broadly acknowledged AI strategies by organizations today, there are different other AI methods. Following is an outline of the absolute most acknowledged ML techniques –

Regulated Learning

These calculations are prepared utilizing marked models, in various situations, as an info where the ideal result is as of now known. A gear, for example, could have information focuses, for example, "F" and "R" where "F" speaks to "fizzled" and "R" speaks to "runs". A learning calculation will get a lot of information guidelines alongside the comparing exact results. The learning calculation will at that point contrast the genuine result and the exact result and banner a mistake, if there is any error. Utilizing various strategies, for example, relapse, arrangement, angle boosting, and forecast, directed learning utilizes various examples to proactively anticipate the estimations of a name on extra unlabeled information. This technique is usually utilized in territories where verifiable information is utilized to anticipate occasions that are probably going to happen later on. For example, envision when a charge card exchange is probably going to be fake or anticipate which protection clients are probably going to record their cases.

Unaided Learning

This technique for ML discovers its application in zones where information has no chronicled marks. Here, the framework won't be furnished with the "right answer" and the calculation ought to distinguish what is being appeared. The fundamental point here is to examine the information and recognize an example and structure inside the accessible informational index. For example, this kind of learning distinguishes client fragments with comparable traits and afterward lets the business to treat them comparatively in showcasing efforts. So also, it can likewise distinguish properties that separate client fragments from each other. Either ways, it is tied in with distinguishing a comparable structure in the accessible informational collection. Furthermore, these calculations can likewise distinguish exceptions in the accessible informational indexes.

Semi-managed Learning

This sort of learning is utilized and applied to a similar sort of situations where regulated learning is pertinent. Notwithstanding, one must note that this strategy utilizes both unlabeled and marked information for preparing. In a perfect world, a little arrangement of named information, alongside a huge volume of unlabeled information is utilized, as it takes less time, cash and endeavors to get unlabeled information. This sort of AI is regularly utilized with strategies, for example, relapse, characterization and expectation. Organizations that generally think that its difficult to meet the significant expenses related with marked preparing process settle on semi-directed learning.

Support Learning

This is primarily utilized in route, apply autonomy and gaming. Activities that yield the best rewards are distinguished by calculations that utilization experimentation techniques. There are three significant parts in fortification learning, to be specific, the operator, the activities and the earth. The operator right now the chief, the activities are what a specialist does, and nature is whatever a specialist connects with. The primary point right now learning is to choose the activities that amplify the prize, inside a predetermined time. By following a decent arrangement, the specialist can accomplish the objective quicker.

3.5. SYSTEM REQUIREMENTS

DATA REQUIREMENTS

The data are necessary as inputs to the analysis, which is specified based upon the requirements of those directing the analysis or customers (who will use the finished product of the analysis). The general type of entity upon which the data will be collected is referred to as an experimental unit (e.g., a person or population of people). Specific variables regarding a population (e.g., age and income) may be specified and obtained. Data may be numerical or categorical (i.e., a text label for numbers).

Accessories:

1. Documentation packages
MS Word 2007
2. Operating System

Any Operating System or windows

Family ahead of Windows 98, for

Desktop applications and Windows 98

Networking or Win NT/2K for network

Base multi-user system.

HARDWARE REQUIREMENTS

System : Pentium IV 2.4 GHz
Hard Disk : 40 GB
Floppy Drive : 1.44 Mb
Monitor : 15 VGA Colour
Mouse : Wired or Wireless
Ram : 512 Mb

SOFTWARE REQUIREMENTS

Operating system : Windows 10
IDE : anaconda navigator
Coding Language : python

Chapter-4

RESULTS AND DISCUSSION

4.1. RESULTS

Thus, the model explains the clustering algorithm using K-means and the streaming algorithm for detecting the crimes. The crime hotspot detection makes it easy for the police officials to go through the crimes in a particular area. The new method of live streaming brings in large amount of data to be handled using clustering.

4.2. DISCUSSION

Profile distinction is obtained for the crime and it is supported records, there are differing kinds that varies from rough to violent, theft, kidnap cases and other crimes. Each crime profile has some internal seriousness pattern. Determining each pattern of the crimes and calculating percentage of the illegal activities is the main aim of the profile difference and this gives the range and variation of the crimes. To calculate the profile distinction for the various crimes, the basic demand is to require the account that handles the criminal activities and also the corresponding profile variations and the main defendant is hold on within the database system. After analyzing the crime data, some stages are involved in noting of the data. Recorded work is completed by adding the lost data. Unwanted and interrupted data will be erased and are eliminated and this gives a predetermined about the contradictory data. The next primary step is the inclusion of data which is the combination of unlike data after the completion of primary step. This step is also the basic step for collecting the data. The final step in this process to select the data which are required for the clustering. : The data streams are generated by various sources that include video police investigation systems, stock markets, sensor networks, web traffic etc.

The results of information stream mining is that the knowledge which will facilitate in taking intelligent and best choices. the information stream mining methodology analyzes the data that is high-dimensional, quick dynamical and may be in time period.

Chapter-5

CONCLUSION AND FUTURE WORK

In this model concepts using data mining have been researched and resulted. Criminal profile is analyzed and processed using data mining. Using clustering methodologies, clusters are easily evaluated and distributed. The clusters are processed before the data mining technique is approached. The forecast in crime are handled by using this data mining. The summary are evaluated supported similarly for all the potential and similar pairs of crimes using a new corresponding method. Streaming data mining concept is used for rapid data records and the strategy of live data streaming, online streaming media is that viewers can predict which area is dangerous and the details about the whereabouts of the crime are known. The protective privacy that is that the future work which ubiquitous in numerous information science disciplines, as well as however not limited to bioinformatics, homeland security, and monetary analysis.

The Software developed is found to be working efficiently and effectively. It results in regular and timely action against crime reported. It can be observed that the information can be obtained easily and accurately. The Software is made user friendly to the maximum so that any lay man can run the software provided he could access to the system via the login password. It believes that partnership work is highly beneficial to the organization and that partnership work is the way forward to reduce crime and disorder. Hence, we wish to remind that its culture should fully endorse partnership work; we urge to ensure the attention they require. It influences law-breaking considerably with the recent development of data collection and circulation technologies. This would be the future work for our model and this might have several advantages like security and conserving data from unauthorised person.

REFERENCES:

1. Yadav, S., Timbadia, M., Yadav, A., Vishwakarma, R., &Yadav, N. (2017, April). Crime pattern detection, analysis & prediction. In Electronics, Communication and AerospaceTechnology (ICECA), 2017 International conference of (Vol. 1, pp. 225-230). IEEE.
2. Shamsuddin, N.H.M.,Ali,N.A. &Alwee, R. (2017, May). An overview on crime prediction methods. In Student Project Conference (ICT-ISPC), 2017 6th ICT International (pp. 1-5). IEEE.
3. Sivaranjani, S., Sivakumari, S., &Aasha, M. (2016, October). Crime prediction and forecasting in Tamilnadu using clustering approaches. In Emerging Technological Trends (ICETT), International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 09 | Sep 2018 www.irjet.net p-ISSN: 2395-0072 © 2018, IRJET | Impact Factor value: 7.211 | ISO 9001:2008 Certified Journal | Page 1042 International Conference on (pp. 1-6). IEEE.
4. Sathyadevan, S., &Gangadharan, S. (2014, August). Crime analysis and prediction using data mining. In Networks & Soft Computing (ICNSC), 2014 First International Conference on (pp. 406-412). IEEE. [1]
[SEP]
5. Nath, S. V. (2006, December). Crime pattern detection using data mining. In Web intelligence and intelligent agent technology workshops, 2006. wi-iat 2006 workshops. 2006 ieee/wic/acm international conference on (pp. 41-44). IEEE.
6. Zhao, X., & Tang, J. (2017, November). Exploring Transfer Learning for Crime Prediction. In Data Mining Workshops (ICDMW), 2017 IEEE International Conference on (pp. 1158-1159). IEEE.
7. Al Boni, M., & Gerber, M. S. (2016, December). AreaSpecific Crime Prediction Models. In Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on (pp. 671-676). IEEE.
8. Tayebi, M. A., Gla, U., &Brantingham, P. L. (2015, May). Learning where to inspect: Informatics (ISI), 2015 IEEE International Conference on (pp. 25-30). IEEE

APPENDIX

- a. Sample Code
- b. Screen Shots

a. Sample Code

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression #import library
from sklearn.svm import SVR
train_data = pd.read_csv("train.csv") #import dataset
print(train_data.shape)      #print row and columns
target = train_data["Category"].unique() # train category column
print(target)
test_data = pd.read_csv("test.csv") #import testing dataset
print(test_data.head())      #print dataset
print(test_data.shape)      #print rows and columns
data_dict = { }             #convert to integer
count = 1
for data in target:
    data_dict[data] = count
    count+=1
train_data["Category"] = train_data["Category"].replace(data_dict)
data_dict
train_data["Category"]      #train the category column
#Replacing the day of weeks
data_week_dict = {
    "Monday": 1,
    "Tuesday":2,
    "Wednesday":3,
```

```

    "Thursday":4,
    "Friday":5,
    "Saturday":6,
    "Sunday":7
}
train_data["DayOfWeek"] = train_data["DayOfWeek"].replace(data_week_dict)
test_data["DayOfWeek"] = test_data["DayOfWeek"].replace(data_week_dict)
train_data["DayOfWeek"]
test_data["DayOfWeek"]
#District
district = train_data["PdDistrict"].unique()
data_dict_district = {}
count = 1
for data in district:
    data_dict_district[data] = count
    count+=1
train_data["PdDistrict"] = train_data["PdDistrict"].replace(data_dict_district)
test_data["PdDistrict"] = test_data["PdDistrict"].replace(data_dict_district)
train_data["PdDistrict"]
test_data["PdDistrict"]
columns_train = train_data.columns
print(columns_train)
columns_test = test_data.columns
print(columns_test)
cols = columns_train.drop("Resolution") #drop resolution column
print(cols)
train_data_new = train_data[cols] #after drop resolution-> train
print(train_data_new.head())
print(train_data_new.describe())
train_data_new.dtypes #print the datatypes
corr = train_data_new.corr()

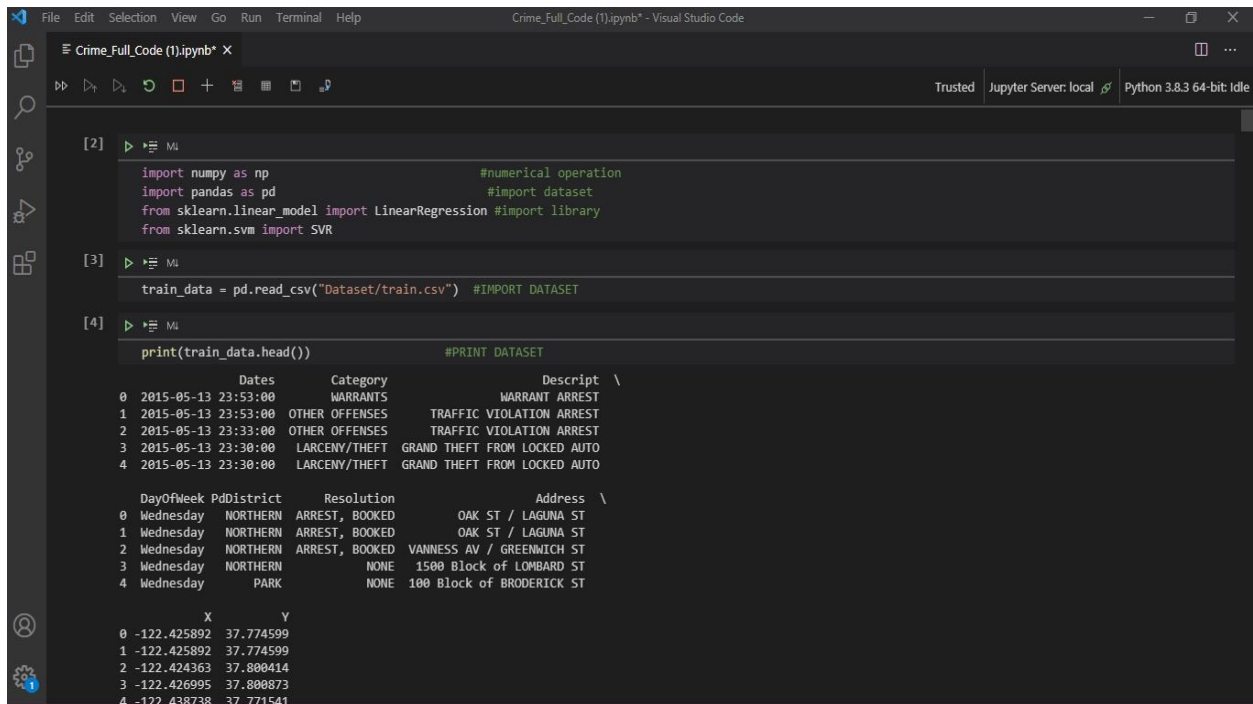
```

```

print(corr["Category"])
skew = train_data_new.skew()
print(skew)
features = ["DayOfWeek", "PdDistrict", "X", "Y"] # future values
X_train = train_data[features]
y_train = train_data["Category"] # going to find category
X_test = test_data[features]
from collections import OrderedDict
data_dict_new = OrderedDict(sorted(data_dict.items()))
print(data_dict_new)
#knn algorithm
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
predictions = knn.predict(X_test)
result_dataframe = pd.DataFrame({
    "Id": test_data["Id"]})
for key,value in data_dict_new.items():
    result_dataframe[key] = 0
count = 0
for item in predictions:
    for key,value in data_dict.items():
        if(value == item):
            result_dataframe[key][count] = 1
    count+=1
result_dataframe.to_csv("submission_knn.csv", index=False)
result_dataframe.to_csv

```

b. Screen Shots



The screenshot shows a Jupyter Notebook in Visual Studio Code with the following code cells:

```
[2] In [ ]: import numpy as np           #numerical operation
import pandas as pd           #import dataset
from sklearn.linear_model import LinearRegression #import library
from sklearn.svm import SVR

[3] In [ ]: train_data = pd.read_csv("Dataset/train.csv") #IMPORT DATASET

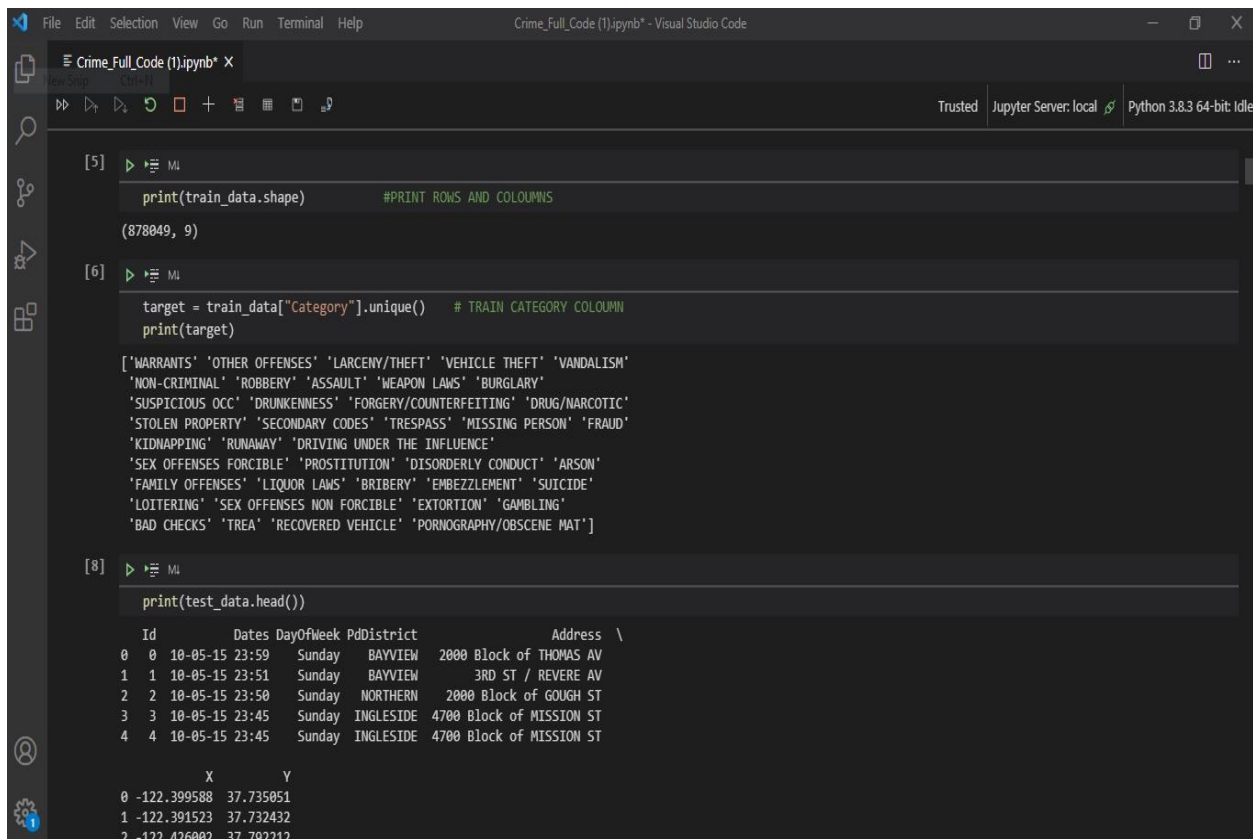
[4] In [ ]: print(train_data.head())           #PRINT DATASET
```

The output of the fourth cell shows the first five rows of the training dataset:

	Dates	Category	Descript \
0	2015-05-13 23:53:00	WARRANTS	WARRANT ARREST
1	2015-05-13 23:53:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST
2	2015-05-13 23:33:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST
3	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO
4	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO

	DayOfWeek	PdDistrict	Resolution	Address \
0	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST
1	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST
2	Wednesday	NORTHERN	ARREST, BOOKED	VANNES AV / GREENWICH ST
3	Wednesday	NORTHERN	NONE	1500 Block of LOMBARD ST
4	Wednesday	PARK	NONE	100 Block of BRODERICK ST

	X	Y
0	-122.425892	37.774599
1	-122.425892	37.774599
2	-122.424363	37.800414
3	-122.426995	37.800873
4	-122.438738	37.771541



The screenshot shows the continuation of the Jupyter Notebook with the following code cells:

```
[5] In [ ]: print(train_data.shape)           #PRINT ROWS AND COLOUWWS

(878049, 9)

[6] In [ ]: target = train_data["Category"].unique() # TRAIN CATEGORY COLOUWWS
print(target)

['WARRANTS' 'OTHER OFFENSES' 'LARCENY/THEFT' 'VEHICLE THEFT' 'VANDALISM'
'NON-CRIMINAL' 'ROBBERY' 'ASSAULT' 'WEAPON LAWS' 'BURGLARY'
'SUSPICIOUS OCC' 'DRUNKENNESS' 'FORGERY/COUNTERFEITING' 'DRUG/NARCOTIC'
'STOLEN PROPERTY' 'SECONDARY CODES' 'TRESPASS' 'MISSING PERSON' 'FRAUD'
'KIDNAPPING' 'RUNAWAY' 'DRIVING UNDER THE INFLUENCE'
'SEX OFFENSES FORCIBLE' 'PROSTITUTION' 'DISORDERLY CONDUCT' 'ARSON'
'FAMILY OFFENSES' 'LIQUOR LAWS' 'BRIBERY' 'EMBEZZLEMENT' 'SUICIDE'
'LOITERING' 'SEX OFFENSES NON FORCIBLE' 'EXTORTION' 'GAMBLING'
'BAD CHECKS' 'TREA' 'RECOVERED VEHICLE' 'PORNOGRAPHY/OBSCENE MAT']

[8] In [ ]: print(test_data.head())
```

The output of the eighth cell shows the first five rows of the test dataset:

	Id	Dates	DayOfWeek	PdDistrict	Address \
0	0	10-05-15 23:59	Sunday	BAYVIEW	2000 Block of THOMAS AV
1	1	10-05-15 23:51	Sunday	BAYVIEW	3RD ST / REVERE AV
2	2	10-05-15 23:50	Sunday	NORTHERN	2000 Block of GOUGH ST
3	3	10-05-15 23:45	Sunday	INGLESIDE	4700 Block of MISSION ST
4	4	10-05-15 23:45	Sunday	INGLESIDE	4700 Block of MISSION ST

	X	Y
0	-122.399588	37.735051
1	-122.391523	37.732432
2	-122.426002	37.792212

```
File Edit Selection View Go Run Terminal Help Crime_Full_Code (1).ipynb* - Visual Studio Code

Crime_Full_Code (1).ipynb* X

[2] In [2]: test_data = pd.read_csv("Dataset/test.csv") #IMPORT TESTING DATASET

NameError                                Traceback (most recent call last)
<ipython-input-2-1a776b1076dc> in <module>
----> 1 test_data = pd.read_csv("Dataset/test.csv") #IMPORT TESTING DATASET

NameError: name 'pd' is not defined

[9] In [9]: print(test_data.shape)

(884262, 7)

[10] In [10]: data_dict = {} #CONVERT TO INTEGER
count = 1
for data in target:
    data_dict[data] = count
    count+=1
train_data["Category"] = train_data["Category"].replace(data_dict)
data_dict

{'WARRANTS': 1,
'OTHER OFFENSES': 2,
'LARCENY/THEFT': 3,
'VEHICLE THEFT': 4,
'VANDALISM': 5,
'NON-CRIMINAL': 6,
'ROBBERY': 7,
'ASSAULT': 8,
```

```
File Edit Selection View Go Run Terminal Help Crime_Full_Code (1).ipynb* - Visual Studio Code

Crime_Full_Code (1).ipynb* X

'WEAPON LAWS': 9,
'BURGLARY': 10,
'SUSPICIOUS OCC': 11,
'DRUNKENNESS': 12,
'FORGERY/COUNTERFEITING': 13,
'DRUG/NARCOTIC': 14,
'STOLEN PROPERTY': 15,
'SECONDARY CODES': 16,
'TRESPASS': 17,
'MISSING PERSON': 18,
'FRAUD': 19,
'KIDNAPPING': 20,
'RUNAWAY': 21,
'DRIVING UNDER THE INFLUENCE': 22,
'SEX OFFENSES FORCIBLE': 23,
'PROSTITUTION': 24,
'DISORDERLY CONDUCT': 25,
'ARSON': 26,
'FAMILY OFFENSES': 27,
'LIQUOR LAWS': 28,
'BRIBERY': 29,
'EMBEZZLEMENT': 30,
'SUICIDE': 31,
'LOITERING': 32,
'SEX OFFENSES NON FORCIBLE': 33,
'EXTORTION': 34,
'GAMBLING': 35,
'BAD CHECKS': 36,
'TREA': 37,
'RECOVERED VEHICLE': 38,
'PORNOGRAPHY/OBSCENE MAT': 39}

[11] In [11]: train_data["Category"] #TRAIN THE CATEGORY COLUMN
```

```

File Edit Selection View Go Run Terminal Help
Crime_Full_Code (1).ipynb - Visual Studio Code

[-] ▶ M4
[-] ▶ M4

[18] ▶ M4
columns_train = train_data.columns #AFTER CHANGING THE VALUES -> TRAINING
print(columns_train)
columns_test = test_data.columns
print(columns_test)

Index(['Dates', 'Category', 'Descript', 'DayOfWeek', 'PdDistrict',
      'Resolution', 'Address', 'X', 'Y'],
      dtype='object')
Index(['Id', 'Dates', 'DayOfWeek', 'PdDistrict', 'Address', 'X', 'Y'], dtype='object')

[19] ▶ M4
cols = columns_train.drop("Resolution") #DROP RESOLUTION COLOUMN
print(cols)

Index(['Dates', 'Category', 'Descript', 'DayOfWeek', 'PdDistrict', 'Address',
      'X', 'Y'],
      dtype='object')

[20] ▶ M4
train_data_new = train_data[cols] #AFTER DROP RESOLUTION-> TRAIN
print(train_data_new.head())

   Dates      Category      Descript  DayOfWeek \
0  2015-05-13 23:53:00         1  WARRANT ARREST         3
1  2015-05-13 23:53:00         2  TRAFFIC VIOLATION ARREST         3

```

```

File Edit Selection View Go Run Terminal Help
Crime_Full_Code (1).ipynb - Visual Studio Code

dtype='object')

[20] ▶ M4
train_data_new = train_data[cols] #AFTER DROP RESOLUTION-> TRAIN
print(train_data_new.head())

   Dates      Category      Descript  DayOfWeek \
0  2015-05-13 23:53:00         1  WARRANT ARREST         3
1  2015-05-13 23:53:00         2  TRAFFIC VIOLATION ARREST         3
2  2015-05-13 23:33:00         2  TRAFFIC VIOLATION ARREST         3
3  2015-05-13 23:30:00         3  GRAND THEFT FROM LOCKED AUTO         3
4  2015-05-13 23:30:00         3  GRAND THEFT FROM LOCKED AUTO         3

PdDistrict      Address      X      Y
0         1  OAK ST / LAGUNA ST -122.425892  37.774599
1         1  OAK ST / LAGUNA ST -122.425892  37.774599
2         1  VANNESS AV / GREENWICH ST -122.424363  37.800414
3         1  1500 Block of LOMBARD ST -122.426995  37.800873
4         2  100 Block of BRODERICK ST -122.438738  37.771541

[21] ▶ M4
print(train_data_new.describe())

   Category      DayOfWeek      PdDistrict      X \
count  878049.000000  878049.000000  878049.000000  878049.000000
mean    7.224975      3.992691      6.037957    -122.422616
std     6.111544      1.972023      3.114045      0.030354
min     1.000000      1.000000      1.000000    -122.513642
25%     3.000000      2.000000      3.000000    -122.432952
50%     5.000000      4.000000      6.000000    -122.416420
75%    10.000000      6.000000      9.000000    -122.406959
max    39.000000      7.000000     10.000000    -120.500000

Y

```

```
File Edit Selection View Go Run Terminal Help
Crime_Full_Code (1).ipynb* - Visual Studio Code

Y
count 878849.000000
mean 37.771020
std 0.456893
min 37.707879
25% 37.752427
50% 37.775421
75% 37.784369
max 90.000000

[22] ▶ M4
train_data_new.dtypes #train the datatypes

Dates      object
Category   int64
Describe   object
DayOfWeek  int64
PdDistrict int64
Address    object
X          float64
Y          float64
dtype: object

[23] ▶ M4
corr = train_data_new.corr() #
print(corr["Category"])

Category      1.000000
DayOfWeek    -0.016263
PdDistrict    0.007643
X             0.000147
Y            -0.005303
Name: Category, dtype: float64

Python 3.8.3 64-bit 0 0 0
Screenshots WhatsApp - Goo... Crime_Full_Cod...
```

```
File Edit Selection View Go Run Terminal Help
Crime_Full_Code (1).ipynb* - Visual Studio Code

[23] ▶ M4
corr = train_data_new.corr() #
print(corr["Category"])

Category      1.000000
DayOfWeek    -0.016263
PdDistrict    0.007643
X             0.000147
Y            -0.005303
Name: Category, dtype: float64

[24] ▶ M4
skew = train_data_new.skew() #
print(skew)

Category      1.662607
DayOfWeek    -0.005572
PdDistrict    -0.232137
X            18.685494
Y           113.984988
dtype: float64

[25] ▶ M4
features = ["DayOfWeek", "PdDistrict", "X", "Y"] # future values
X_train = train_data[features]
y_train = train_data["Category"] # going to find category
X_test = test_data[features]

[5] ▶ M4
#knn algorithm
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=5)
```

```

Crime_Full_Code (1).ipynb* - Visual Studio Code
Crime_Full_Code (1).ipynb* X
Trusted | Jupyter Server: local | Python 3.8.3 64-bit: Idle

[29] In [29]:
for key,value in data_dict_new.items():
    result_dataframe[key] = 0
count = 0
for item in predictions:
    for key,value in data_dict.items():
        if(value == item):
            result_dataframe[key][count] = 1
            count+=1
result_dataframe.to_csv("submission_knn.csv", index=False)
result_dataframe.to_csv

<bound method NDFrame.to_csv of
0 0 0 1 0 0 0 0 0 0
1 1 0 0 0 0 0 0 0 0
2 2 0 0 0 0 0 0 0 0
3 3 0 0 0 0 0 0 0 0
4 4 0 0 0 0 0 0 0 0
5 5 0 0 0 0 0 0 0 0
6 6 0 0 0 0 0 0 0 0
7 7 0 0 0 0 0 0 0 0
8 8 0 0 0 0 0 0 0 0
9 9 0 0 0 0 0 0 0 0
10 10 0 0 0 0 0 0 0 0
11 11 0 1 0 0 0 0 0 0
12 12 0 1 0 0 0 0 0 0
13 13 0 0 0 0 0 0 0 0
14 14 0 0 0 0 0 0 0 0
15 15 0 0 0 0 0 0 0 0
16 16 0 0 0 0 0 0 0 0
17 17 0 1 0 0 0 0 0 0
18 18 0 0 0 0 0 0 0 0
19 19 0 0 0 0 0 0 0 0

```

```

Crime_Full_Code (1).ipynb* - Visual Studio Code
Crime_Full_Code (1).ipynb* X
Trusted | Jupyter Server: local | Python 3.8.3 64-bit: Idle

DRUNKENNESS ... SEX OFFENSES NON FORCIBLE STOLEN PROPERTY SUICIDE \
0 0 ... 0 0 0 0
1 0 ... 0 0 0 0
2 0 ... 0 0 0 0
3 0 ... 0 0 0 0
4 0 ... 0 0 0 0
5 0 ... 0 0 0 0
6 0 ... 0 0 0 0
7 0 ... 0 0 0 0
8 0 ... 0 0 0 0
9 0 ... 0 0 0 0
10 0 ... 0 0 0 0
11 0 ... 0 0 0 0
12 0 ... 0 0 0 0
13 0 ... 0 0 0 0
14 0 ... 0 0 0 0
15 0 ... 0 0 0 0
16 0 ... 0 0 0 0
17 0 ... 0 0 0 0
18 0 ... 0 0 0 0
19 0 ... 0 0 0 0
20 0 ... 0 0 0 0
21 0 ... 0 0 0 0
22 0 ... 0 0 0 0
23 0 ... 0 0 0 0
24 0 ... 0 0 0 0
25 0 ... 0 0 0 0
26 0 ... 0 0 0 0
27 0 ... 0 0 0 0
28 0 ... 0 0 0 0
29 0 ... 0 0 0 0
... ..
884232 0 ... 0 0 0 0
884233 0 ... 0 0 0 0

```