IJIRMET
ISSN (Online): 2456-0448
International Journal of Innovative Research in Management, Engineering, and Technology
Vol. 7, Issue 3, March 2022

Crime Data Analytic And Prediction Using Machine Learning Algorithm

[1] Ms.R.Srividya M.E,[2] Mr.K.Chairmadurai M.E,[3] M.Sandhiya

[1] Assistant professor, Department of Computer Science and Engineering, Adhiparasakthi Engineering College, Melmaruvathur, India.

Assistant professor, Department of Computer Science and Engineering, Adhiparasakthi Engineering College, Melmaruvathur, India.

Department of Computer Science and Engineering, Adhiparasakthi Engineering College, Melmaruvathur, India.

Abstract: ***Crimes damage any society each socially and economically. enforcement bodiesface varied challenges whereas making an attempt to stop crimes. we have a tendencyto propose acriminaloffense information Analytics Platform (CDAP) to help enforcement bodies to perform descriptive, predictive, and prescriptive analysis on crime information. CDAP contains a standard design wherever every part is constructed one by one from the others. CDAP conjointly supports plugins sanctionative future feature expansions. The platform will ingest any crime dataset that has needed the specified the desired} attributes to map the dataset to attributes required by the platform. It will then analyze them, train models, so visualize information. CDAP conjointly combines census information with crime information to realize a additional comprehensive crime analysis and its impact on society. Moreover, with the mix of census information and crime information, CDAP provides method reengineering steps to optimize resource allocations of police forces. we have a tendency to demonstrate the utility of the platform by visualizing spacial and temporal relationships in a very set of real-world crime datasets. prognosticative capabilities of the platform ar incontestable by predicting crime classes, that a machine learning approach is employed. To construct a model area theorem, Random Forest Classifier, and Multi-layer Perceptron Network classification algorithms ar provided. Identification of optimized police district boundaries and allocating patrol beats ar accustomed demonstrate the prescriptive analytics capabilities of the tool. Heuristic-based clump approach was taken to outline police district boundaries in a very manner that the known districts have equitable population distribution with compact shapes. The ensuing districts ar then evaluated on the difference of population and also the compactness exploitation the Gini constant and Isoperimetric Quotient. Another heuristic-based approach was taken to outline new police patrol beats to be optimized on equitable work distribution, compactness, and minimizing reaction time for brand spanking new police patrol beats.***

Keywords: CDAP, Nave Bayesian, Random Forest, Multi-layer perceptron.

1. INTRODUCTION

GENERAL

Crimes square measure a social nuisance and it's an immediate result onsociety.Governments pay countless cash through enforcement agencies to undertake and stop crimes from happening. Today, several enforcement bodies have giant volumes of knowledge associated with crimes, which require to be processed to show into helpful info. Crime knowledge square measure complicated as a result of they need several dimensions and in several formats, e.g., most of them contain string records and narrative records. because of this diversity, it's troublesome to mine them employing a shelf, applied mathematics, and machine learning knowledge analytics tools. it's the first reason for the dearth of a general platform for crime data processing. whereas there square measure some propitiatory platforms to predict and analyze crime knowledge, they're targeted solely on bound areas of crimes, not protractile, associated don't offer an API to integrate with alternative tools. Moreover, identical tool can not be used for the analysis and well as designing like patrol beads and district boundaries.

MOTIVATION

High or enlarged crime levels build communities decline, as crimes scale back house costs, neighborhood satisfaction, and therefore the need to maneuver negatively. to scale back and stop crimes it's necessary to spot the explanations behind crimes, predict crimes, and bring down solutions. because of giant volumes knowledge|of knowledge of information} and therefore the variety of algorithms required to be applied to crime data, it's unrealistic to try to to manual analysis. Therefore, it's necessary to own a platform that's capable of applying any formula needed to try to to a descriptive, predictive, and prescriptive analysis on an oversized volume of crime information. Through those 3 methodologies, law-enforcement authorities are ready to take appropriate actions to forestall the crimes. Moreover, by predicting the extremely probably targets to be attacked, throughout a selected amount and specific geographical location, police are ready to determine higher ways in which to deploy the restricted resources and conjointly to seek out and fix the issues

resulting in crimes. many applications square measure already developed for crime analysis. Most of those tools square measure developed to assist the police to spot totally different crime patterns and even to predict criminal activities. they're complicated software system that desires tons of coaching before use. coming up with a tool that's simple to use with bottom coaching would facilitate law-enforcing bodies all round the world to scale back crimes.

PROBLEM STATEMENT

The analysis downside that this project tries to deal with may be expressed as follows:
How to develop a software system platform to conduct descriptive, predictive, and prescriptive analysis of various crime data?
Descriptive analyzing focuses on distinctive spatial-temporal relationships with crime information. prophetical analytics ways square measure primarily used for predicting the class of a criminal offense which will be occurred somewhere at a given time. to attain it, the system integrates Census information with the crime information and feeds it to machine learning algorithms. Prescriptive instrument, suggests method re-engineering steps portion police resources optimally assuming to scale back crimes and impact to the overall public.
ANALYSIS OBJECTIVES
The analysis objectives of this project square measure as follows:
•Develop a platform which will be accustomed analyze crime information mistreatment descriptive and prophetical information analytics techniques.
• Using the projected platform analyze the spatial and temporal (time of day, day of week, and seasons) relationships in crime information.
• Suggest appropriate method reengineering steps and resource allocations supported the spatial and temporal relationships. as an example,
o Identify new police district boundaries mistreatment Heuristic-Based plane figure clump methodology.
o Identify intelligent patrol routes which will mix crime information and spatial dimensions mistreatment Voronoi Tessellations.
• Analyze the link between crime information and census information.

II LITERATURE REVIEW

A. Crimes and Impact on the Society
Author: Deepak Singh, Bhavana Narain
Year:2021
A crime may be outlined as any action or omission that violates a law, which ends up in a very penalisation. typically what constitutes a criminal offense depends on the govt bodies and laws that square measure existing in those places. to grasp the character of crimes, one needs to perceive not solely its Spatio-temporal dimensions, however conjointly the character of the crime, the victim-offender relationship, the role of guardians, and therefore the history of comparable incidents. notwithstanding the explanations why crimes happen, they place a strain on the communities, towns, and cities. Usual financial prices related to them embrace the price of policing crime and prosecuting people who commit crimes. Non monetary prices accommodates social prices, wherever they have an effect on the standard of life, psychological state, and physical security of individuals living in those areas. Crimes square measure a social nuisance and having the ability to unravel them quicker is extremely necessary and can acquire itself.
B. Sociology Theories
AUTHOR: John and David
YEAR:2019
According to John and David, theories of crimes may be divided into 2 classes specifically, people who request to clarify the event of criminal offenders and people that request to clarify the event of criminal events. sociology has been primarily developed through theories and analysis on offenders. solely recently it's begun to clarify the crimes instead of the criminalness of individuals concerned in them. sociology consists of the many theories that designate however and why some offenders act within the means they are doing. Following are some theories that explain how places are associated with crimes.
1. Rational Choice suggests that offenders will select targets and define means to achieve their goals in a manner that can be explained. Further, it can be explained as that human actions are based on rational decisions, that is they are informed by probable consequences of that action.
2. Routine Activity Theory This theory explains the occurrence of crimes as the result of several circumstances. Namely, a motivated offender, a desirable target, target, an offender must be at the same place at the same time, and lastly absent of other types of controllers intimate handlers, guardians, and place managers.
3. Crime Pattern Theory This theory combines the above two theories and goes on to say that how targets come to the attention of offenders is influenced by the distribution of crime events over time, space, and among targets. An offender will come to know of criminal opportunities while engaging in their day-to-day legitimate work. So given offender will only know about a subset of available targets. The concept of place is essential to crime pattern theory . Having an understanding of criminology theories is essential to try and create crime analysis tools or platforms using modern technologies. Having an understanding of criminology theories is essential

to try and create crime analysis tools or platforms using modern technologies.

C. CRIME ANALYSIS

Crime analysis is a difficult task, as it requires both collection and analysis of large volumes of data. For example, Brown states that Richmond city in the USA has approximately 100,000 criminal records per year. Given the data volume and need to apply different algorithmic techniques forbids manual analysis. Whereas an automated analysis of such a rich data set could identify complex crime patterns and assist in solving crimes faster. Data mining techniques can be used in law and enforcement for crime data analysis, criminal career analysis, bank fraud analysis, and analysis of other critical problems.Some of the traditional data mining techniques are association analysis, classification and prediction, cluster analysis, and outlier analysis, which identify patterns in structured data. Using criminology theories along with modern technology would help to identify crime patterns quickly and efficiently. To simplify the workload a crime data analytic platform could be used which would help in simplifying the process while driving more accurate and insightful conclusions and predictions.

III SYSTEM ANALYSIS

EXISTING SYSTEM

Several applications have been already developed for crime analysis. Most of these tools are developed to help the police forces to identify different crime patterns and even to predict criminal activities. Recent applications were developed by aiming at adopting data mining techniques. Next, we discuss some of the key solutions .

A. Copy Link

Chen et al.describe COPLINK as integrated data and information management atmosphere making an attempt to manage the huge quantity of knowledge on criminals. it's been developed at the University of Arizona's AI work together with the metropolis and Phoenix Police departments. the most aim of this project is to develop data and information management systems, technologies, and methodologies applicable for capturing, accessing, analyzing, visualizing, and sharing law enforcement-related data in social and structure contexts. COPLINK consists of 2 main components particularly, COPLINK connects and COPLNK find. COPLINK connect is accountable for sharing information from completely different police departments, whereas COPLINK find is employed to uncover completely different crime associations that exist in police databases. find is usually involved with making associations and linkages among numerous aspects of against the law. It uses applied math techniques like co-occurrence analysis and clump functions to weight relationships between all attainable pairs of ideas. However, the subsequent drawbacks are often known during this system.

•COPLINK is advanced and needs user coaching.

• Although the system will determine linkages among specific ideas residing in associate existing info, it doesn't support data processing.

• The current version of COPLINK doesn't support temporal reasoning or image.

B. Recap

Regional Crime Analysis Program (ReCAP) could be a system designed to assist native police forces through crime analysis and bar. It works with the handgun 2000 records management system. The system is sort of recent and is predicated on Windows ninety five and NT, and works solely with a neighborhood space Network (LAN). this technique has 3 main components, particularly a info, geographic data system, and data processing tools. It provides the subsequent key functions.

•Hot Sheet provides a outline of the foremost necessary crime activity of the region.

• Summary Reports These reports tally specific crime occurrences over a user-defined time and space.

Map-oriented Searches offer a GIS

• display of the realm beside the premeditated criminal activity. The crimes are often displayed on the map in step with the sort of crime, time/date of crime, location, suspect description, weapons concerned, etc.

• Time Charting Patterns developing over time of the day, day of the week, etc., area unit premeditated. Indeterminate crimes area unit premeditated statistically mistreatment kernel density estimation.

• Cluster Analysis Uses algorithms like k-means and nearest neighbor to perform clump, to spot statistically vital groupings of crimes in a district.

• Detailed Inter-Modular looking Helps in police investigation links between vehicles, suspects, warrants, etc.

C. Crime Prediction Model

Another resolution is Ozguls Crime Prediction Model (CPM), that predicts offenders of terrorist events supported location, date, and procedure attributes. It uses each resolved and unresolved crime data, learning from the attributes of every crime. These crimes area unit clustered in step with the attributes. As associate example, equally placed crime clusters. These clusters contain crimes in step with specific attributes. Similarity scores area unit measured for every crime and total similarity between 2 crimes is measured mistreatment geometrician distance. mistreatment these values equally behaving resolved and unresolved crimes area unit place into a lot of correct clusters. CPM appearance for perpetrators of crimes, with the idea that the bulk of crimes during a single cluster were committed by identical single wrongdoer cluster. This platform focuses chiefly on terrorist activities and teams.

D. Rationalize Police Patrol Beats Mistreatment Voronoi Tessellation

Many connected works target optimizing offered resources to cut back crimes. Suresh et. al.demonstrates a way to use Voronoi Tessellations to divide a given police jurisdiction into a collection of patrol beats for the equitable employment. As an indication of idea, they need used a sample information set from the state capital department of local government. completely different weights were allotted to crimes per their severity. Then Police beats were distributed per crime sorts, crime data, crime teams, and earth science. As future work, authors have planned to increase the Voronoi tessellations to fulfill completely different necessities of the police. One such thought is to rationalize the boundaries of police beats by taking road networks and physical landscapes like rivers under consideration. this will be additionally extended by considering information like bad person residencies for higher police work. However, the authors haven't enforced a framework or any quite implementation on the theories mentioned apart from an indication of idea.

E. Optimal Choice of Police Patrol Beats

• Mitchell planned a model for choose patrol beats supported a heuristic approach. It takes some quantity of assumptions to require under consideration whereas proposing the model. Among those assumptions the incident distribution, over each area and time, is assumed, which a distance live or metric between the centers of every fractional monetary unit is assumed and additionally the closest offered unit responds to a decision. A heuristic-based approach was utilized in model building wherever the subsequent heuristic was used.

• minimize ∑minimum (i, j)

W(i,j) is that the matrix of befittingly weighted distances. The assumptions of the model need that W(i, j) be a distance matrix with the (i, j)-th part representing the weighted travel or different distance from the i-th location to the j-th location. the target is then to settle on a set of k rows of W in such a fashion on minimize the total of the column

• minimums, wherever every column minimum is chosen solely from among the selected set of k rows. The heuristic formula Mitchel planned has 2 phases. within the 1st section, k locations ar elect in some fashion. within the next improvement section, the formula seeks to boost on locations elect within the 1st section, by ordered substitution of the locations elect. the method is perennial for every of the locations not within the allocation till no improvement is formed once a whole cycle. There ar many different platforms and models represented in many papers concerning crime information analysis. Revathy and Satheesh mentioned many different solutions like Self Organizing Map (SOM) that links sexual offenders of sexual attacks. every of the on top of platforms and solutions assists law-enforcement bodies to analyse and establish completely different crime patterns. one in all the most reasons for developing completely different platforms to research crime information is that the Brobdingnagian volume of information that's required to be analyzed. This task has become not possible to try and do manually. loads of analysis is finished on ways in which to spot crime patterns mistreatment completely different recursive ways like cluster utilized in Crime Prediction Model. Taking these offered resources into usage is extremely vital. Providing one platform that's capable of mistreatment {different|totally completely different|completely different} techniques utilized in different platforms is very important to research and establish crime patterns. By mistreatment prognosticative models, authorities will establish which type of crimes may occur most in an exceedingly given amount around that areas. distinguishing these details is extremely vital for various enforcement authorities to create selections on a way to minimize crime. A platform having the ability to research completely different crime patterns descriptively would facilitate to spot patterns in crime and a few platforms already give this facility having the ability to predict the sort of crimes that might occur in an exceedingly} given space at a given time is additionally very helpful and prediction is employed by crime prediction model to spot terrorists concerned in terrorist activity. A platform having the ability to produce prescriptive analysis on ways in which to reduce crimes may stop crimes from happening. this might additionally facilitate enforcement authorities to create use of their restricted resources within the handiest method. once longing these solutions, it's clear that these platforms ar specific to a given task. it's terribly helpful to possess all the preceding analytical techniques in one platform. that's a platform which will be extended to produce descriptive, predictive, and prescriptive analysis of crime information.

TOOLS, ALGORITHMS, AND INFRASTRUCTURE

TOOLS

A. Apache Storm

Apache Storm could be a free and ASCII text file distributed period of time computation system. Storm processes massive volumes of high-speed information in period of time. it's extraordinarily quick andmay method over 1,000,000 records per second per node on a cluster of modest size. Storm on Hadoop YARN (Yet Another Resource Negotiator) is powerful for machine learning functions and primarily for period of time analysis. a number of the utilization cases of Storm ar period of time analytics, on-line machine learning, continuous computation, distributed RPC, ETL, and more. Apache Storm operates on a continual stream of information that isn't aiming to happen in our use case for crime analysis. Apache Spark performs data-parallel computations whereas Storm performs Task-Parallel computations.

B. Apache Spark

Apache Spark could be a all-purpose cluster computing engine that's thought of in no time and reliable. it's AN ASCII text file platform for large-scale processing that's appropriate for unvaried machine learning tasks. Apache Spark provides Application Programming Interfaces (APIs) in programming languages like Java, Python, and Scala. Spark provides generality by powering a stack of libraries which incorporates SQL and information Frames, MLlib for machine learning, GraphX, and Spark Streaming. These libraries ar designed upon the Apache Spark core and that they will seamlessly mix within the same application. Spark runs on Hadoop, Mesos, standalone, or perhaps within the cloud. Spark supports in-memory computing that has enabled it to question information a lot of quicker compared to disk-based engines like Hadoop. Apache Spark web site offers statistics that prove this reality. It shows that Spark runs programs up to 100x quicker than Hadoop MapReduce in memory and 10x quicker on disk. Spark will analyze massive volumes of crime information, that is needed in our framework. Reasons to settle on Spark :

• Spark uses Resilient Distributed Datasets (RDDs) that permits storing information on memory and dogging it as per the wants. It permits an enormous increase in instruction execution job performance.

•Spark permits caching information in memory, that is helpful within the case of unvaried algorithms that ar employed in machine learning. it's additionally necessary as a result of it's terribly helpful for interactive data processing that is needed by the planned platform.

• It provides a feature to affix informationsets across multiple disparate data sources.

• When examination the time taken for the k-means formula on a dataset of various sizes K-means victimization Spark forever gave far better performance than K-means victimization MapReduce.

For Machine Learning functions Apache Spark provides MLlib, a distributed machine learning library. It consists of the many quick and ascendable implementations of ordinary learning algorithms for common learning settings together with classification, regression, cooperative filtering, clustering, and spatiality reduction. It additionally provides some underlying statistics, algebra, and optimisation primitives. MLlib includes Java, Scala, and Python Apis. due to Spark Integration, this library will use core functionalities for functions like information cleanup and featurization. Spark SQL provides information integration practicality. in comparison performance-wise with Apache Spark, Apache driver v0.9 on Hadoop MapReduce was a lot of slower primarily because of MapReduce programming overhead and lack of support for unvaried computation excluding all this, Apache Spark has intensive documentation and an oversized and active community. Considering all the higher than factors we have a tendency to selected Apache Spark as a tool for our project.

ALGORITHMS

Following are the algorithms which can be used in the proposed solution.

A. Clustering Algorithms

Several papers have mentioned several obtainable algorithms that may be accustomed analyze crime information. Shyam monitor lizard Nath says in his paper that 100 percent of criminals commit five hundredth of the crimes. data processing includes a higher influence in fields like Law and social control for crime issues, crime information analysis, criminal career analysis, bank frauds, and alternative important issues. Following area unit some common clump algorithms that may be used for crime information analysis. Shyam's paper goes on to counsel that the clump technique may be a higher approach than the other supervised techniques like classification since crimes vary in nature wide and crime information usually contains many unresolved crimes. Also, the character of crimes changes over time, thus to spot newer and unknown patterns within the future, clump techniques work higher. K-MEANS clump formula. This formula is especially accustomed partition the clusters supported their mean. As a primary step variety of objects area unit classified and nominal as k clusters. K numbers of objects area unit ab initio elite because the cluster centers. although these objects area unit allotted supported cluster center. Then cluster means that area unit updated once more. This formula is employed as a base for many of the opposite clump algorithms.AK-MOD formula. This is a clump formula consisting of 2 phases. within the initial section, attributes area unit weighted. Weights of the attributes area unit calculated mistreatment the knowledge Gain Ratio(IGR) for every attribute. The attribute with the best worth is taken because the decisive attribute. within the second section (clustering) initial, the quantity of clusters k and also the initial mode of every cluster area unit found. Then distance for each mode and its nearest mode area unit calculated. After that, every cluster mode is updated. This method keeps going until all modes don't seem to be updated once more. EXPECTATION-MAXIMIZATION formula. This is Associate in Nursing extension of the k-means clump formula. it's accustomed calculate parameter estimates for every cluster. Weights of attributes area unit measured in likelihood distribution and every object is to be clustered supported the weights. to live parameter estimates 2 steps area unit followed.

• Expectation Step: during this step, for every object of clusters the likelihood of cluster membership of object x(i) is calculated.

• Maximization Step: Re-estimate/refine model parameters mistreatment estimation from the first step.

•Classification Algorithms: Classification is taken into account a supervised prediction technique. it's been employed in several domains like prognostication, health care, medical, financial, etc. 2 completely different classification algorithms area unit thought-about. they're specifically call Tree and Naive Bayesian. Naive Bayesian is taken into account an efficient formula to resolve classification tasks. a call tree may be a usually used prognostic model and it additionally follows a supervised learning approach. because the name suggests it forms a tree-like structure and every node represents a take a look at on attribute worth. Leaves represent categories that prognostic models for classification. Branches represent conjunctions of options. This formula applies a top-down approach. The gain in entropy

is employed to guide the formula for the creation of nodes. There area unit some execs and cons of each algorithms. Naive Bayesian needs a shorter coaching time and it's a quick analysis. it's additional appropriate for real-world issues. However, for complicated classification issues call tree is additional suited. It produces affordable and explainable classification trees. This paper suggests that the choice Tree has higher accuracy and preciseness over Naive Bayesian.

B. Decision Tree

Decision tree learning could be a technique unremarkably utilized in data processing and machine learning tasks, classification, and regression. call trees break down a dataset into smaller subsets whereas at a similar time Associate in Nursing association tree is incrementally developed. the ultimate tree has call nodes and leaf nodes. call nodes have 2 or additional branches whereas the leaf node represents classification or call. call trees will handle each categorical and numerical knowledge. The core rule for building a call tree is named ID3 that employs a top-down, greedy search through the area of attainable branches with no backtracking. every partition is chosen avariciously by choosing the most effective split from a group of attainable splits, to maximise the data gain at a tree node. data gain is the distinction between the parent node impurity and also the weighted total of the two-child node impurities. Impurity are often measured victimisation Entropy.

Cp − fi log fi equation calculates the entropy. Here, C is that the variety of distinctive labels, fi is that the frequency of label i at a node. benefits of employing a call tree rule ar

•It is easy to grasp and interpret.

•This technique needs little knowledge preparation.

•Decision trees will handle each numerical and categorical knowledge.

•Decision trees perform well with giant datasets. It are often wont to analyze giant datasets among an inexpensive time victimisation normal computing resources.

C. Random Forest Classification

Random forests ar one amongst the foremost flourishing machine learning models for classification and regression. Random forests rule doesn't overfit. it's achieved by combining many choices trees. One will run as several trees PRN. Random forests rule is taken into account quick. the fundamental rule of Random forests trains a group of call trees one by one, therefore the coaching are often exhausted parallel. The rule injects randomness into the coaching method to form every call tree completely different from the opposite. Combining predictions from every tree reduces the variance of predictions, rising the performance of check knowledge. Following ar a number of the options of Random forests classification.

•Random forests handle categorical options and may even be multi-class classification.

•It will capture non-linearity and have interactions.

•This rule runs expeditiously on giant databases.

•It will handle thousands of variables while not variable deletion.

•The Random Forest rule is effective for estimating missing knowledge and maintains accuracy even once giant components of information ar missing.

D. Multi-Layer Perceptrons (MLPs)

Multi-Layer Perceptrons (MLPs) ar classified as a sort of Artificial Neural Network. The computation is performed employing a set of the many easy units with weighted connections between them. Multi-layer perceptron consists of some layers. they're specifically,

• Input layer: this can be the bottom-most layer that takes input from a dataset. it's the exposed a part of the network.

Hidden layers: One or additional layers that ar

• not directly exposed to the input.

• The output layer takes the output from the ultimate hidden layer. It outputs a price or vector of values that correspond to the format needed for the matter.

E. Frequent Pattern- Growth (Fp-Growth)

In this rule, the primary step is to calculate item frequencies and establish frequent things. there's a tree- structure known as Frequent-Pattern tree used for succeeding step. A frequent pattern tree could be a tree structure outlined as below. It consists of 1 root tagged null, a group of item prefix subtrees because the youngsters of the basis, and a frequent-item header table. every node within the item prefix subtree consists of 3 fields: item-name, count, and node-link. every item within the frequent-item

header table consists of 2 fields, specifically item-name and head of node-link, that purpose to the primary node within the FP-tree carrying the item name. FP-growth rule uses this structure to encrypt transactions while not generating candidate sets expressly, that is taken into account a fashionable method. once this method, frequent itemsets are often extracted from the FP-tree.

There ar benefits to victimisation the FP-Growth rule.

ISSN (Online): 2456-0448
International Journal of Innovative Research in Management, Engineering, and Technology
Vol. 7, Issue 3, March 2022

• It has solely 2 passes over the dataset

• It compresses the dataset.

• FP-Growth rule is far quicker than Apriori.

ARCHITECTURES
LAMBDA ARCHITECTURE

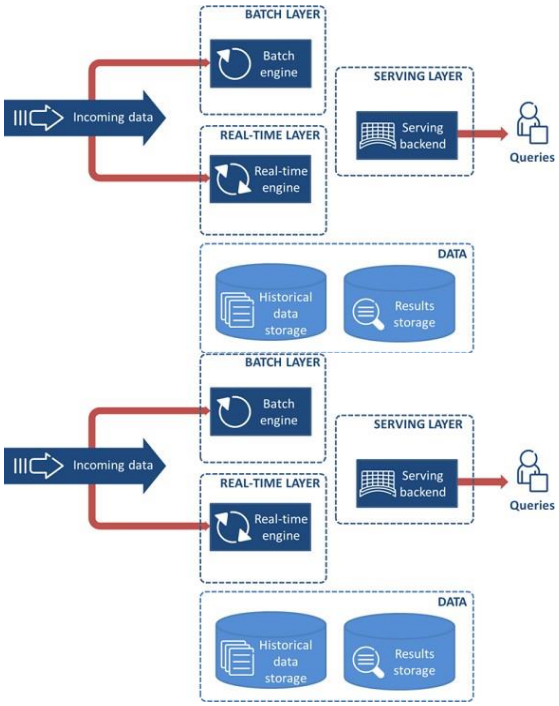


Figure 3.1 Lambda Architecture

Lambda design that is shown in Figure 3.1, may be a generic, scalable, and fault-tolerant processing design. it's a mix of speed layer, batch layer, and serving layer. All the information entered into the system is sent to each the batch layer and speed layer for process. information sent to the batch layer is deposited within the master information set and processed sporadically with a delay and regarded as complete as a result of it retrieved all the historical information. exploitation that historical information instruction execution element builds machine learning models. information sent to the speed layer is processed that information in period and generates period views. however these period views don't seem to be thought of complete. The serving layer depends on the utilization case of the applying it takes action in keeping with the items happening within the batch layer and repair layer. Any incoming question is answered by merging results from batch views and period views. Applying a fancy design like lambda design to the crime information analyticplatform that doesn't need period processing is debatable Crime records area unit unbroken as a dataset that contains info concerning many domains collected over years. thus Crime information doesn't would like stream to the system. they will be fed to the system as historical information sets and it doesn't need period processing.

KAPPA ARCHITECTURE

ISSN (Online): 2456-0448
International Journal of Innovative Research in Management, Engineering, and Technology
Vol. 7, Issue 3, March 2022

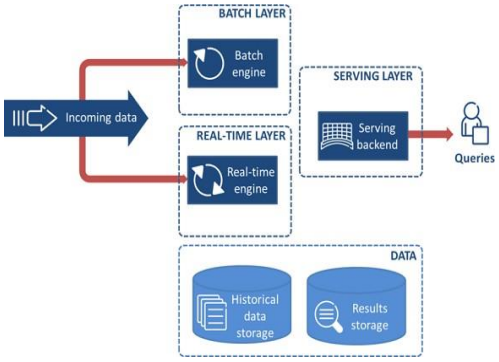


Figure 3.2 Kappa Architecture

Kappa design could be a simplification of Lambda design. A letter design system is sort of a Lambda design system with the instruction execution system removed. to switch instruction execution, information is solely fed through the streaming system quickly. letter design is targeted on a unified Log, all the streaming information will be stirred as a unified log that means one readying wherever each event collected is immutable , ordered, and therefore the current state of the event are going to be modified solely by appending a replacement event. As shown in Figure three.2, all the process of the event are going to be performed within the input streams and persisted as time period views. To support human fault-tolerant, the events also are persisted in storage like HDFS if the info is aged out of the unified logs. In time period to employ a happening begin a second instance of the duty that starts process from the start of the event and directs the output to a unique table, once the reprocessing job is fixed purpose the applying to the new table and drop the recent table. This continues within the streaming layer if one thing has to be reprocessed while not having the Batch Layer. The advantage of this design is, the developer has to employ only the code is modified. If the modified code doesn't work fine he will roll back to the recent output table. Also, it's potential to mirror the Franz Kafka topics to HDFS storage for long-run use.

3.3 PROPOSED DESIGN

The platform that goes to be developed directly targets the crime domain. It will analyze crime knowledge in 3 alternative ways, specifically Descriptive, prescriptive, and prognosticative analytics. the foremost vital a part of this platform is that it's designed to be scalable to support differing kinds of crime knowledge analysis. completely different user necessities are often achieved through developing straightforward plug-ins to the system and scaling the platform. Descriptive analyser uses each quantitative and qualitative knowledge at the side of analytical techniques. Qualitative knowledge and analytical techniques ask non-numerical knowledge, also because the examination and interpretation of observations to get underlying meanings and patterns of relationships. The descriptive analyser provides relationships between crimes and identifies the pattern of crimes and temporal and spacial relationships between crimes. It conjointly provides a applied mathematics outline of a given knowledge set. Prescriptive analyser tries to spot the rationale behind the crimes and provides suggestions to avoid or scale back the crimes. It will determine vital factors associated with the crimes committed. Through plugins, users will manipulate and extend the platform for specific desires. during this platform, prognosticative analytics ways ar in the main used for predicting the class of against the law that may be occurred somewhere at a given time. Predict crime class system integrates population and race knowledge consequently to the given crime knowledge set. victimisation machine learning techniques will predict the class of crime that may occur. Any user or a enforcement body, UN agency contains a crime dataset will use this feature to know the severity of the crime that will come about, and as a result, may take necessary steps to allot resources effectively. the subsequent sections describe the design and completely different modules enforced within the resolution. Features provided by the platform. The following options ar provided by the CDAP.

1.Redraw economical police jurisdiction boundaries.

2. Question knowledge within the crime dataset.

3.Draw economical police patrol beats supported the crime distribution.

4.Predict crime classes for a given crime scene.

ISSN (Online): 2456-0448
International Journal of Innovative Research in Management, Engineering, and Technology
Vol. 7, Issue 3, March 2022

5. transfer against the law dataset and population dataset.

6. Preprocess the uploaded crime dataset.

IV SYSTEM REQUIREMENTS

FUNCTIONAL REQUIREMENTS

The Functional requirements specify which input should be given to the model and which output should be produced by the model. Each functional requirement should specify a detailed description of all data and their sources.

NON-FUNCTIONAL REQUIREMENTS

These are the quality constraints that the system must satisfy according to the project contract. The priority or extent to which these factors are implemented varies from one project to another. They are also called non-behavioral requirements.

HARDWARE REQUIREMENTS

- Processor : Any Processor above 500 MHz.
- Ram : 4 GB
- Hard Disk : 4 GB
- Input device : Standard Keyboard and Mouse.
- Output device : VGA and High Resolution Monitor.

SOFTWARE REQUIREMENTS

- Operating System: Windows 7 or higher
- Programming: Python 3.6 and related libraries
- Software : Anaconda Prompt(Anaconda3)

V PROJECT DESCRIPTION

Depicts the high-level design of the planned Inquisitors Crime knowledge Analytics Platform (CDAP). knowledge Receiver is employed to supply knowledge to the platform and also the knowledge persistence unit is employed to store knowledge employed by the system and trained models. The preprocessor is employed to preprocess data received by the information Receiver that successively fed into subsequent layer. It consists of 2 core modules, applied mathematics analyser and Machine Learner. exploitation those 2 modules, the Descriptive analyser, Prescriptive analyser, and prognosticative analyser are enforced. Also, there area unit a couple of different elements that give an important contribution to the CDPA. The functionalities that area unit provided by prescriptive, descriptive, ANd prognosticative analyzers area unit exposed to the user through an API. The results that area unit provided by CDAP consistent with the user's requests will be visualised exploitation the visualiser part.

SYSTEM ARCHITECTURE

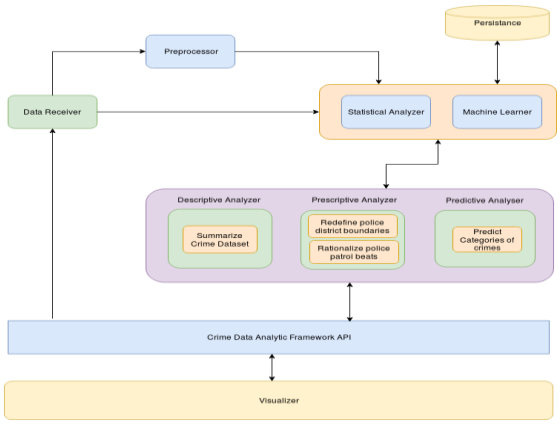


Figure 5.1 High-level architecture of the proposed platform

A. Preprocessor

The module Preprocessor that is employed by all different modules consists of the options and functionalities that square measure required to preprocess the information before feeding them into different modules like Machine Learner, applied math

analyser. among this module, functionalities required to preprocess the input file are enforced. when receiving the CSV record from the user-specified location, the information extracted from that file are processed in 3 stages.

• knowledge Cleaning: Data is clean through processes like filling in missing values, smoothing the blatant knowledge, or breakdown the inconsistencies within the knowledge.

•Data Integration: Data with totally different representations square measure place along and conflicts among the information square measure resolved.

•Knowledge Transformation: Data is normalized, aggregated, and generalized.

These are enforced exploitation Spark among Java. The processed knowledgeset are given to the opposite module as Spark data frames.

B. Applied Math Analyser

The applied math analyser provides a basic data point of the information set the feeds to the framework. It provides API to induce easy statistics like mean, variance median, column statistics to complicated statistics like FP growth algorithmic program primarily based frequent itemsets. The applied math analyser module provides AN API to perform basic applied math analysis on the crime dataset that is provided by the user. It permits the user to induce the subsequent analytical outcomes.

•Overall outline of the dataset

•Column wise statistics: Mean, Variance, Median.

•Complex statistics: Frequent itemsets supported FP growth algorithmic program.

• Functions and performance prams

C. Machine Learner and Predictive Analyzer

In this platform, prophetical analytics strategies area unit primarily used for predicting the class of a criminal offense which will be occurred somewhere at a given time. To integrate prophetical analytics options, it's necessary to own a machine learning element likewise. There area unit many attainable approaches to integrate the machine learning element with the platform. Since Apache Spark along side its Machine Learning Library R provides a fashionable facility for machine learning and data processing techniques, victimization Apache Spark R may be a stronger selection. Using R inside the platform may be enforced in many alternative ways. One is victimization the java library that provides the power to decision and use R functions from java programs. Another one is victimization the SparkR library that is provided by spark itself. analysis through with Associate in Nursing FBI information set has steered regression toward the mean because the best formula for predictions. we've got enforced the Machine Learning element that is constructed on high of Apache Spark MlLib and created during a thanks to hide the quality of spark Machine learning algorithms. Since e-crime information contains a ton of matter information it generates large inconsistencies and errors once about to feed into the cc element. This module provides the subsequent functionalities.

->Random Forest Classification
->Multilayer Perceptron Classification
->Decision Tree Classification

D. Descriptive Analyser

Descriptive analyser use each quantitative and qualitative information and analytical techniques. Qualitative information and analytical techniques talk to non-numerical information likewise because the examination and interpretation of observations to find underlying meanings and patterns of relationships. this can be most common of field analysis, content analysis, and historical analysis. Quantitative information area unit information primarily in numerical or categorical format. The chemical analysis consists of manipulations of observations make a case for} and explain the phenomena that those observations mirror and is primarily applied mathematics. Descriptive Crime analysis employs each forms of information and techniques reckoning on the analytical and sensible would like. as an example, crime information may be utilized in varied ways in which, each quantitatively and qualitatively. the knowledge like date, time, location, and sort of crime is quantitative in this statistics may be wont to analyze these variables. with the exception of the spatial distribution of crimes, this considers the temporal relationship between the crimes, geographical state of affairs, Population density, and atmospheric condition to relinquish a a lot of elaborated description of the crime.

E. Prescriptive Analyzer

Urban area unit as are divided into many police districts sometimes supported demographic and geographic parameters. this can be primarily done to divide the work and to deal with the problems that embrace homicides, burglaries, and other forms of crimes

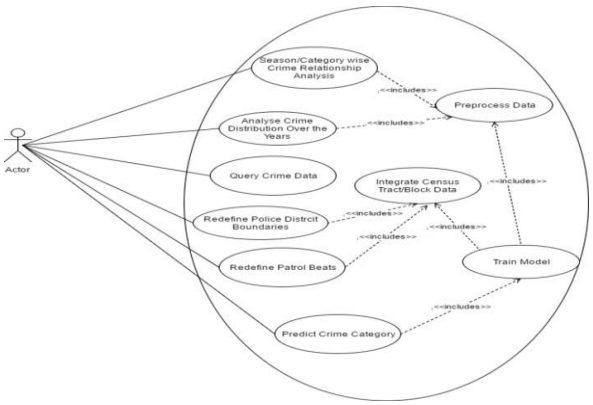
a lot of effectively by dividing the resources on the market for the police forces during a lot of economical means. as an example, point of entry is split into ten police districts. Most of the police districts that exist nowadays are fashioned a few years agone and with changes happening over time and also the movement of population, we've got to question whether or not the prevailing police districts give for the aim they were created. As a part of the answer, the Inquisitors Crime information Analytic platform will redraw police districts' boundaries regarding the distribution of population during a a lot of uniform manner. The platform provides the potential to redistrict a given country for any quantity of attainable police districts in line with the population distribution. The second objective of the sub-system Prescriptive analyser is shown in three.1, System will rationalize police patrol beats within redrawn police districts. The system takes Road network distance rather than ancient manhattan distance to calculate weighted crime distance between points. Also, the system takes geographical constraints under consideration by about to the tiniest geographical census level on the market. Inquisitor Crime information Analytic platform uses a heuristic-based approach to rationalize police patrol beats inefficient means. to boost the practicality of police patrol beats and create them simpler, work distribution among crimes are going to be equal and supply terribly low police patrol interval.

F. Application Programming Interface (API)

Each of the core modules prophetical analyser, Prescriptive analyser, Descriptive analyser, and Preprocessor use facade style patterns to cover the complexness of the module and supply necessary functionalities through a simplified interface. victimization Crime knowledge Analytics Framework API provides all the functionalities of the framework through associate degree API. each kinds of functionalities that embrace within the facades and not enclosed within the facades may be gained through the Crime knowledge Analytics Framework API. for instance Preprocessor Facade within the preprocessor module solely provides the functionalities to handle missing knowledge by deleting the complete row. however through the API, the user will choose regardless of the missing price handling choice enforced within the preprocessor module in keeping with his/her preference (like predicting the missing price of replacement victimization the frequent item then on). 5.1.7 beholder The beholder element allows the user to interactively analyze the dataset and therefore the results provided by the prophetical, prescriptive, and applied math analyser. This element provides knowledge visual image through models like histograms, heat maps, tables, pin maps. the info needed for the visual image is retrieved by connecting with the prescriptive, prophetical and applied math analyzers through the provided API.

UMLDIAGRAM

UML stands for Unified Modeling Language. UML could be a standardized all-purpose modeling language within the field of object-oriented computer code engineering. the quality is managed and was created by, the thing Management cluster. The goal is for UML to become a typical language for making models of object-oriented laptop computer code. In its current type, UML is comprised of 2 major components: a Meta-model and a notation. within the future, some type of methodology or method might



also be intercalary to; or related to, UML.

USECASE READ

The crime knowledge Analysis platform has been designed to handle specific use cases. Figure 5.2 shows use cases of the CADP.

Figure 5.2 Use case diagram for CDAP

DEVELOPMENT VIEW

ISSN (Online): 2456-0448
International Journal of Innovative Research in Management, Engineering, and Technology
Vol. 7, Issue 3, March 2022

Figure 5.3 explains the whole system from the developer's perspective. Six main components are loosely coupled with each other.

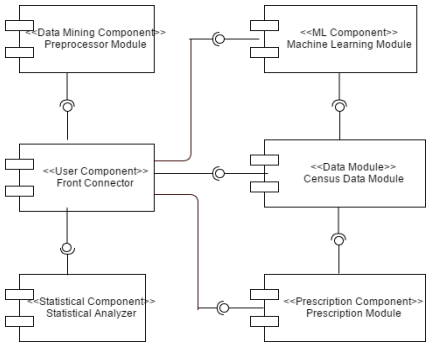


Figure 5.3development diagram for CDAP

PREPROCESSOR MODULE

Preprocessor Module consists of the options and functionalities that ar required to preprocess the info before feeding them into alternative modules like Machine Learner and applied mathematics analyser. inside this module, functionalities required to preprocess the computer file are enforced. Front connecter module. The front connecter provides API to the user by concealing the complexness of the Crime knowledge Analytic Framework. Users will feed knowledge into the framework and acquire results back mistreatment the front connecter module.

STATISTICAL ANALYSER MODULE

The applied mathematics analyzer provides a basic data point of the info set the feeds to the framework. It provides API to induce straightforward statistics like mean, variance median, column statistics to complicated statistics like FP growth rule based mostly frequent item sets. The applied mathematics analyzer module provides associate API to perform basic applied mathematics analysis on the crime dataset that is provided by the user. Machine learning module The Machine Learning element is made on prime of Apache Spark MlLib and created in a very thanks to hide the complexness of spark Machine learning algorithms. Census knowledge module. This module keeps varied forms of census knowledge like census block knowledge, census tract knowledge race knowledge.

KNOWLEDGE DESCRIPTION

This platform is predicated on the Apache Spark engine. Therefore, inside the framework within the platform, knowledge are unbroken user shouldn't got to bear in mind of the way to use spark, those spark knowledge structures are hidden inside some and processed as knowledge structures utilized by Apache Spark like knowledge Frame, Dataset. however since the new custom knowledge structures that permits the user to use the platform simply while not information concerning Apache Spark. Basic knowledge flow inside the platform is shown in Figure 5.4.

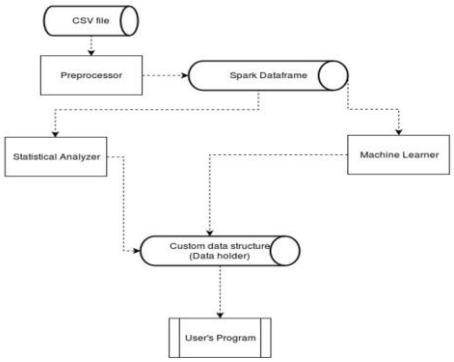


Figure 5.4: Data flow within CDAP

VI SAMPLE CODE

```
from flask import Flask, request
from flask_restful import Resource, Api
from sqlalchemy import create_engine
```

```
from json import dumps
from flask.json import jsonify
from flask_cors import CORS
import json
from flasgger import Swagger
# install
# pip install flask flask-jsonpify flask-sqlalchemy flask-restful flask_cors flasgger
import sys, json, io, time
import pandas as pd
import numpy as np
from keras.models import model_from_json
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder
from keras.utils import np_utils
app = Flask(__name__)
cors = CORS(app, resources={r"/api/*": {"origins": "*"}})
api = Api(app)
Swagger(app)
@app.route('/api/records', methods=['POST'])
def records_index(month=12, day=15, hour=6):
"""
This API predict crime happens index
Call this api passing occurrence month, occurrence day, occurrence hour --- tags: - dasboard crime prediction parameters: - name: body
in: body
schema:
id: Date and Period
required:
- month
- day
- hour
properties:
month:
type: integer
description: month
default: 12
day:
type: integer
description: day
default: 15
hour:
type: integer
description: hour
default: 6
responses:
200:
description: The required result is available
500:
description: Error!
"""
jsonObj = request.get_json()
month = jsonObj.get('month')
print(month)
day = jsonObj.get('day')
hour = jsonObj.get('hour')
return records_prediction_handler(month, day, hour)
```

```
@app.route('/api/location', methods=['POST'])
def address_index(address = " 1 Yonge St, Ontario", month=12, day=15, hour=6):
"""
properties:
month:
type: string
description: address
default: 1 yonge street Ontario
month:
type: integer
description: month
default: 12
day:
type: integer
description: day
default: 15
hour
type: integer
description: hour
default: 6
responses:
200:
description: The required result is available
500:
description: Error!
"""
jsonObj = request.get_json()
month = jsonObj.get('month')
print(month)
day = jsonObj.get('day')
hour = jsonObj.get('hour')
return records_prediction_handler(month, day, hour)
result = NB[NB.Hood_ID == value+1]
#print(result.Neighbourhood)
return result.Neighbourhood
def getperiod(x):
if x=='0':
period='Night'
elif x=='6':
period='Morning'
elif x=='12':
risk='Medium'
else:
risk='High'
return risk
def loaded_model(model):
#loding model
# load json and create model
json_file = open('/'+model+'_structure.json', 'r')
loaded_model_json = json_file.read()
json_file.close()
loaded_model = model_from_json(loaded_model_json)

n = len(poly)
inside =False
```

```
p1x,p1y = poly[0]
for i in range(n+1):
p2x,p2y = poly[i % n]
if y > min(p1y,p2y):
if y <= max(p1y,p2y):
if x <= max(p1x,p2x):
if p1y != p2y:
xinters = (y-p1y)*(p2x-p1x)/(p2y-p1y)+p1x
if p1x == p2x or x <= xinters:
inside = not inside
p1x,p1y = p2x,p2y
return inside
def coordinate_to_neighbourhood(lat,lon,neigReverseID, city_geojson):
flag=-1
for index in range(1,141):
if(point_inside_polygon(lat, lon, city_geojson.features[index1]["geometry"]["coordinates"][0] )): topoMapID=index
neigReverseID = pd.read_json('/neighbours_reverse.json')
city_geojson = pd.read_json('/toronto_geojson.json')
app.run(threaded=False, port=5000, host='0.0.0.0')
```

VII SNAPSHOTS

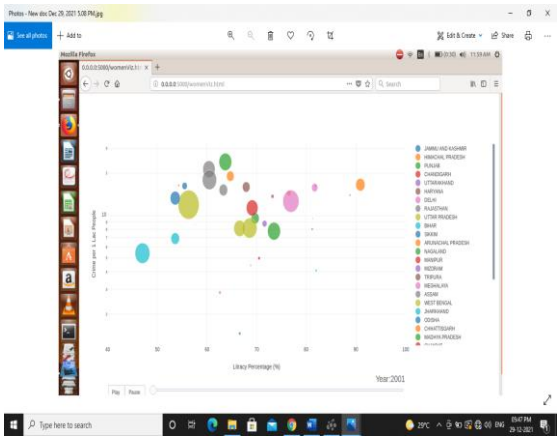
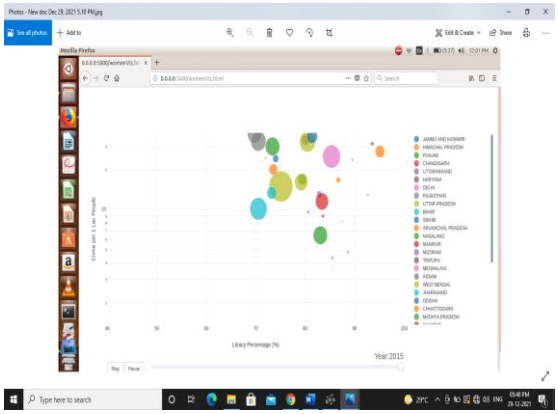


Figure 7.1 In year of 2001 crime prediction



ISSN (Online): 2456-0448
International Journal of Innovative Research in Management, Engineering, and Technology
Vol. 7, Issue 3, March 2022

Figure 7.2 In year of 2007 crime prediction

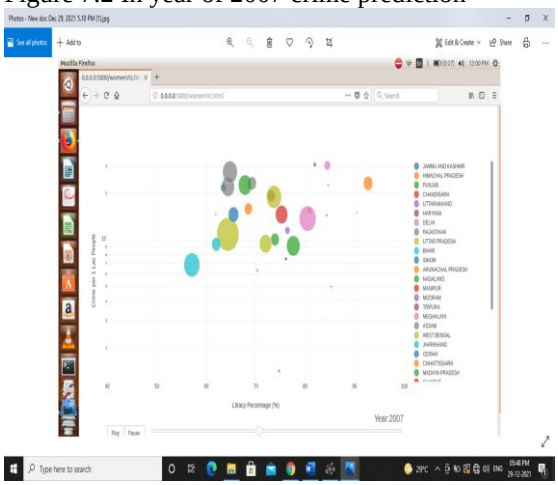


Figure 7.3 In year of 2015 crime prediction

CONCLUSION

We have proposed a heuristic-based, clustering method to divide a given police district into optimal patrol beats based on crime and census data. We use past crime data, their severity, census data, and geographic data to determine patrol beats and optimum seed points (i.e., optimum positioning of the police patrol car). The calculated Gini coefficient for new workload distribution in police patrol beats for 10 police districts was 0.036 and it suggests that the workload variance among districts is low. Also, as the compactness measure, we have used Isoperimetric Quotient and we got an average value of 0.71 for each police patrol beat it suggests that the generated police patrol beats are more compact, and therefore it can minimize the police response time as expected. It can be proved by observing the average time taken for a police patrol car to arrive at any point inside the police patrol beats which is 25.22 seconds. As a case study, we have applied our algorithm to redefine the police district boundaries of San Francisco County. To measure the inequality of population distribution of newly created police districts we have used the Gini Coefficient. The calculated Gini Coefficient for new population distribution was 0.02468 and it suggests that the population variance among districts is low. Also, as the compactness measure, we have used Isoperimetric Quotient and we got an average value of 0.4957 for each district and which suggests the generated districts are compact. Hence, we can conclude that our algorithm has succeeded in the process of redistricting police district boundaries by generating a more compact and equally distributed population.

REFERENCES

[1] S.K.Lodha and A.K.Verma, "Spatio-temporal visualization of urban crimes on a gis grid," 8th ACM Intl. symposium on Advances in Geographic Information Systems, pp. 174–179, 2000.
[2] J. Forgeat. (2015) Data processing architectures lambda and kappa. [Online]. Available: https://www.ericsson.com/research-blog/data-knowledge/ data-processing-architectures-lambda-and-kappa/
[3] C. Yu, M. W. Ward, M. Morabito, and W. Ding, "Crime forecasting using data mining techniques," 11th IEEE Intl. Conf. on Data Mining Workshops, pp. 779–786, 2011.
[4] A. T. Murray, I. McGuffog, J. S. Western, , and P. Mullins, "Exploratory spatial data analysis techniques for examining urban crime implications for evaluating treatment," British Journal of criminology, vol. 41, no. 2, pp. 309–329, 2001.
[5] R. Krishnamurthy and J. S. Kumar, "Survey of data mining techniques on crime data analysis," International Journal of Data Mining Techniques and Applications, vol. 1, no. 2, pp. 117–120, 2012.
[6] D. E. Brown, "The regional crime analysis program (recap): a framework for mining data to catch criminals," IEEE Intl. Conf. on Systems, Man, and Cybernetics, vol. 3, pp. 2848–2853, 1998.
[7] H. Chen, D. Zeng, H. Atabakhsh, W. Wyzga, and J. Schroeder, "Coplink: managing law enforcement data and knowledge," Communications of the ACM, vol. 46, no. 1, pp. 28–34, 2003.

[8] A. Verma, R. Ramyaa, S. Marru, Y. Fan, and R. Singh, "Rationalizing police patrol beats using voronoi tessellations," IEEE Intl. Conf. on Intelligence and Security Informatics (ISI), pp. 165–167, 2010.
[9] J. E. Eck and D. L. Weisburd, "Crime places in crime theory," Crime and place: Crime prevention studies, vol. 4, 2015.

[10] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," Computer, vol. 37, no. 4, pp. 50–56, 2004.
[11] P. S. Mitchell, "Optimal selection of police patrol beats," Criminal Law and Criminology, vol. 63, p. 577, 1972
[12] A.S.Foundation. (2015) Apache storm. [Online]. Available: http://storm.apache. org/
[13] S.Gopalani and R.Arora, "Comparing apache spark and map reduce with performance analysis using k-means," International Journal of Computer Applications, vol. 113, no. 1, pp. 8–11, 2015.

", Knowledge-Based Systems (222) 107020.Year-2021.