**M. S. Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR nagar, MSRIT post, Bangalore-54

A Dissertation Report on

Analysis of Crime Patterns in India

Submitted by

MURTAZA HAIDRI 1MS12CS062

URJA N 1MS12CS125

SHUBHAM AGRAWAL 1MS12CS105

SHRUTHI N 1MS12CS104

*in partial fulfillment for the award of the degree of*

# *Bachelor of Engineering in Computer Science & Engineering*

>



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

[www.msrit.edu](http://www.msrit.edu), **May 2015**

**ACKNOWLEDGEMENT**

First and foremost we would like to thank Mrs. Sowmya Bj (Associate professor), Department of Computer Science and Engineering, M. S. Ramaiah Institute of technology, Bangalore, who was also our supervisor, for their consistent guidance on each and every step of our project and also our HOD Dr. K. G. Srinivasa (Professor). Without their push and motivation, this project would have not been completed. Their continuous support and help made this project feasible. Without her support we would not have been able to excel in our project. Lastly, we would like to thanks all our friends and teammates for believing in each other and our capabilities.

# ABSTRACT

A major challenge facing all law-enforcement and intelligence gathering organizations is accurately and efficiently analyzing the growing volumes of crime data. There has been an enormous increase in the crime in the recent past. The concern about national security has increased significantly since the 26/11 attacks. Crimes are a social nuisance and cost our society dearly in several ways. Here we look at use of missing value, clustering algorithm and Anomalies detection for a data mining approach to help predict the crimes patterns and speed up the process of solving crime. We will look at MV algorithm, DBScan and PAM outlier detection algorithm with some enhancements to aid in the process of filling the missing value and identification of crime patterns. We applied these techniques to real crime data. We can use semi-supervised learning technique here for knowledge discovery from the crime records and to help increase the predictive accuracy.

#### DECLARATION

We hereby declare that the project work entitled **“Analysis of Crime Patterns in India”** submitted to M. S. Ramaiah Institute of Technology, Bangalore, is a record of an original work done by us under the careful guidance of  **Dr. & Prof. K. G. Srinivas,** Head, Dept of Computer Science & Engineering, and this project work is submitted in the partial fulfillment of the requirements for the award of the degree of Bachelor of Engineering in Computer Science & Engineering. The results embodied in this project have not been submitted to any other University or Institute for the award of any degree or diploma.

#### CONTENTS

***Declaration***

***Acknowledgements***

***Abstract***

***List of Figures***

***List of Tables***

1 **INTRODUCTION**

* 1. General Introduction……………….
  2. Statement of the Problem…………..
  3. Objectives of the project……………
  4. Project deliverables……………
  5. Current Scope………………………
  6. Future Scope……………………….

1. **PROJECT ORGANIZATION**
   1. Software Process Models
   2. Roles and Responsibilities
2. **LITERATURE SURVEY**

3.1….Introduction

3.2…Main Body

3.3 Conclusion of Survey

1. **SOFTWARE REQUIREMENT SPECIFICATIONS**

5.1 Product Overview

5.2 External Interface Requirements

5.2.1 User Interfaces

5.2.2 Hardware Interfaces

5.2.3 Software Interfaces

5.2.4 Communication Interfaces

5.3 Functional Requirements

5.3.1 Functional Requirement 1.1

:

5.3.n Functional Requirement 1.n

5.4 Software System Attributes

5.4.1 Reliability

5.4.2 Availability

5.4.3 Security

5.4.4 Portability

5.4.5 Maintainability

5.4.6 Performance

5.5 Performance Requirements

5.6 Database Requirement

5.7 Design Constraints

5.8 Other Requirements

1. **DESIGN**
   1. Introduction
   2. Architecture Design
   3. Graphical User Interface
   4. Class Diagram and Classes (Represent Inheritance, Aggregation and Association)
   5. Sequence Diagram
   6. Data flow diagram
   7. Metric calculation
2. **IMPLEMENTATION**
   1. Tools Introduction
   2. Technology Introduction
   3. Overall view of the project in terms of implementation
   4. Explanation of Algorithm and how it is been implemented
   5. Information about the implementation of Modules
3. **TESTING**

7.1 Results and Snapshots

1. **CONCLUSION & SCOPE FOR FUTURE WORK**
2. **REFERENCES**

**LIST OF FIGURES**

Fig 1. System Architecture

Fig 2. Flow Diagram

**INTRODUCTION**

Crime is a major issue where the top priority has given by our government. Criminology is an area that focuses the scientific study of crime and criminal behavior and law enforcement and is a process that aims to identify crime characteristics. It is one of the most important fields where the application of data mining techniques can produce important results. Crime analysis, a part of criminology, is a task that includes exploring and detecting crimes and their relationships with criminals.

Law enforcement agencies like that of police today are faced with large volume of data that must be processed and transformed into useful information. The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Identifying crime characteristics is the first step for developing further analysis. The knowledge that is gained from data mining approaches is a very useful tool which can help and support police forces [9]. According to [13], solving crimes is a complex task that requires human intelligence and experience and data mining is a technique that can assist them with crime detection problems. The idea here is to try to capture years of human experience into computer models via data mining.

Now-a-days, the criminals are becoming technologically sophisticated in committing crimes [1]. Therefore, police needs such a crime analysis tool to catch criminals and to remain ahead in the eternal race between the criminals and the law enforcement. The police should use the current

Technologies [4] to give themselves the much-needed edge. Availability of relevant and timely information is of utmost necessity in conducting of daily business and activities by the police, particularly in crime investigation and detection of criminals. Police organizations everywhere have been handling a large amount of such information and huge volume of records. There is an urgent need to analyze the increasing number of crimes as approximately 17 lakhs Indian Penal Code (IPC) crime, and 38 lakhs local and Special Law crimes per year.

An ideal crime analysis tool should be able to identify crime patterns quickly and in an efficient manner for future crime pattern detection and action. However, in the present scenario, the following major challenges are encountered.

·Crime information volume has been increased.

·Problem of identifying techniques that can accurately and efficiently analyze this growing volumes of crime data.

·Different methods and structures used for recording crime data.

·The data available is inconsistent and are incomplete thus making the task of formal analysis a far more difficult.

·Investigation of the crime takes longer duration due to complexity of issues

All the above challenges motivated this research work to focus on providing solutions that can enhance the process of crime analysis for identifying and reducing crime in India. The main aim of this research work consist of developing analytical data mining methods that can systematically address the complex problem related to various form of crime. Thus, the main focus is to develop a crime analysis tool that assists the police in:

·To perform crime analysis to detect crime patterns.

·Provide information to formulate strategies for crime prevention and reduction.

·Identify and analyze common crime patterns to reduce further occurrences of similar incidence.

The present research work proposes the use of an amalgamation of data mining techniques that are linked with a common aim of developing such a crime analysis tool. For this purpose, the following specific objectives were formulated.

·To develop a data cleaning algorithm that cleans the crime dataset.

·To explore and enhance clustering algorithms to identify crime patterns from historical data.

·To explore and enhance classification algorithms to predict future crime behavior based on previous crime trends.

**PROBLEM STATEMENT**

Crime Analysis in Karnataka.

In this project, we look at predictability of yearly totals of crime in districts of Karnataka, given data detailing characteristics of ethnicity, education, employment, population, environment etc. for every district.

**OBJECTIVES OF THE PROJECT**

Through the course of human civilization, law and order has been a constantly pressing issue, dealt by monarchs and democracies in varying ways. In this project, we look at predictability of yearly totals of crime (concentrating on burglaries) in districts of Karnataka, given census data detailing characteristics of ethnicity, education, employment, population, transport, environment (natural and built) for every neighborhood. One hopes that this may be of use to social scientists, criminologists, law enforcement agencies and city planners. The main focus is to develop a crime analysis tool that assists the police in :

* To perform crime analysis to detect crime patterns.
* Provide information to formulate strategies for crime prevention and reduction.
* Identify and analyze common crime patterns to reduce further occurrences of similar incidence.

**PROJECT ORGANIZATION**

**1. Software Process Models**

**Requirement Analysis & Definition**: All requirements of the system which has to be developed are collected in this step. Like in other process models requirements are split up in functional requirements and constraints which the system has to fulfil. Requirements have to be collected by analysing the needs of the end user(s) and checking them for validity and the possibility to implement them. The aim is to generate a Requirements Specification Document which is used as an input for the next phase of the model.The datasets and the tools and techniques were collected, studied and their pros and cons were analysed in this phase.  
  
**System Design**: The system has to be properly designed before any implementation is started. This involves an architectural design which defines and describes the main blocks and components of the system, their interfaces and interactions. The three stages pre-processing of data, processing of data and the visualizations and predictions were defined..

**Software Design:** Based on the system architecture which defines the main software blocks the software design will break them further down into code modules. The interfaces and interactions of the modules are described, as well as their functional contents. All necessary system states like startup, shutdown, error conditions and diagnostic modes have to be considered and the activity and behaviour of the software has to be defined. The output of this phase is a Software Design Document which is the base of the following implementation work. The algorithms for clustering and classification, k-means, DBScan and MV algorithm were defined and their behaviour was documented.   
  
**Coding:** Based on the software design document the work is aiming to set up the defined modules or units and actual coding is started. The system is first developed in smaller portions called units. They are able to stand alone from an functional aspect and are integrated later on to form the complete software package.   
  
**Software Integration & Verification:** Each unit is developed independently and can be tested for its functionality. This is the so called Unit Testing. It simply verifies if the modules or units to check if they meet their specifications. This involves functional tests at the interfaces of the modules, but also more detailed tests which consider the inner structure of the software modules. During integration the units which are developed and tested for their functionalities are brought together. The modules are integrated into a complete system and tested to check if all modules cooperate as expected.   
  
**System Validation:** After successfully integration including the related tests the complete system has to be tested against its initial requirements. This will include the original hardware and environment, whereas the previous integration and testing phase may still be performed in a different environment or on a test bench. The system is tested repeatedly with different data sets and the performance is observed.

**2. Roles & Responsibilities**

**Murtaza Haidri**

* Literature survey involving reading various research papers and various implementation techniques.
* Collection of data from different sources.
* Learning of the tools and techniques, R and SAP LUMIRA.
* Data cleaning: Removal of unwanted data and processing it.
* Installation of the tools R, RStudio and SAP Lumira for the clustering, classification and visualization of data.
* Implementation of the k-means algorithm
* Implementation of the analysis involving DBScan and MV algorithm.
* Report making
* Prepare SRS
* Linking of the csv, excel and such other files with data obtained from the sources.
* Created visualiztions of the processed data using SAP Lumira using histograms, Paretto charts, pie charts, heat maps and the likes.

**Shruthi N**

* Literature survey involving reading various research papers and various implementation techniques.
* Collection of data from different sources.
* Report making
* Proof reading the documents for errors and omissions.
* Preperation of the documentation of Software Process Models.
* Testing the project
* Documentation of the results
* Linking of the csv, excel and such other files with data obtained from the sources.

**Shubham Agrawal**

* Literature survey involving reading various research papers and various implementation techniques.
* Collection of data from different sources.
* Learning of the tools and techniques, R and SAP LUMIRA.
* Data cleaning: Removal of unwanted data and processing it.
* Implementation of the k-means algorithm
* Implementation of the analysis involving DBScan and MV algorithm.
* Report making
* Prepare SRS
* Design the system architecture
* Design the flow of the project
* Linking of the csv, excel and such other files with data obtained from the sources.
* Created visualiztions of the processed data using SAP Lumira using histograms, Paretto charts, pie charts, heat maps and the likes.

**Urja N**

* Literature survey involving reading various research papers and various implementation techniques.
* Collection of data from different sources.
* Learning of the tools and techniques, R and SAP LUMIRA.
* Data cleaning: Removal of unwanted data and processing it.
* Installation of the tools R, RStudio and SAP Lumira for the clustering, classification and visualization of data.
* Design the system architecture
* Design the flow of the project
* Implementation using the shinyapps and Quandl libraries of R.
* Implementation of the k-means algorithm
* Implementation of the analysis involving DBScan and MV algorithm.
* Linking of the csv, excel and such other files with data obtained from the sources.
* Created visualizations of the processed data using SAP Lumira using histograms, pareto charts, heat maps, pie charts amongst the others.
* Report and documentation

**LITERATURE SURVEY**

**1. Introduction**

As an introduction to crime analysis, this section provides the definition of crime and crime analysis as a general concept as well as definitions of six types of crime analysis. These definitions are meant to enhance the understanding of crime analysis and to help create commonly understood terminology.

**1.1. Crime definition**

Crime is a multifaceted concept that can be defined in legal and non-legal sense. From a legal point of view, it refers to breaches of the criminal laws that govern particular geographic areas (jurisdictions) and are aimed at protecting the lives, property and rights of citizens within those jurisdictions. Most of the crimes with which the criminal justice system is concerned involve breaches of State/Territory legislation that cover most offences relating to persons (for example, murder and sexual assault), property (for example, theft and property damage) and regulation (for example, traffic violations). Commonwealth legislation relates primarily to matters such as trade and commerce, importa- tion/exportation, taxation, defence and external affairs . [3]

Non-legal point of view would define crime as acts that violate socially accepted rules of human ethi- cal or moral behaviour. As the moral principles that underpin the notion of crime are subject to grad- ual change over time, the types of behaviour defined by the legal system as criminal may also change. Examples of behaviours that have been de-criminalized in some jurisdictions include prostitution, abortion, attempted suicide and homosexual intercourse. Other behaviours, such as tax evasion or credit card fraud, have been criminalized over time .

The distinction between these two points of view is important. While criminal justice agencies' crime data will be based upon the relevant legal definition of crime, data collected via victimisation surveys may be based upon an individual's interpretation of crime, rather than upon existing legal definitions. This potential disparity is more likely to affect the less serious end of the 'crime' spectrum. [4][5]

**2. Types of crime analysis**

The following are five types of analysis of crime. As you will see, each contains characteristics of crime analysis in general, but each is specific in the type of data and analysis used as well as in its purpose.

Not all law enforcement analysts do the same type of work. There are several different types of policy analysis. Crime analysts spend varying percentages of their time on the following types of analysis : [5]

**Tactical Crime Analysis:**

This is day-to-day crime analysis, looking for series, patterns, sprees, hot spots, and hot dots immedi- ately affecting the jurisdiction. Tactical crime analysis also focuses on specific information about each crime such as method of entry, point of entry, suspects actions, type of victim, type of weapon used, as well as the date, time, location, and type of location. Field information such as suspicious activity calls for service, criminal trespass warnings, and persons with scars, marks, or tattoos col- lected by officers is also considered in the analysis. Used for:

-Day to day

-For series, patterns, sprees, hot spots

**Strategic Crime Analysis:**

The study of crime and law enforcement information integrated with socio-demographic and spatial factors to determine long term “patterns” of activity, to assist in problem solving, as well as to re- search and evaluate responses and procedures. Used for:

-Identify unusual activity levels by time or location.

-Forecasting potential crime events/concentration.

**Administrative/Academic Crime Analysis:**

The study of crime and law enforcement information integrated with socio-demographic and spatial factors to determine long term “patterns” of activity, to assist in problem solving, as well as to re- search and evaluate responses and procedures. Used for:

-Reports or statistical summaries for grant funding, commanders & public

-Policy implications beyond law enforcement agency.

**Operations Analysis:**

The study of policing practices is regards patrol and resource allocation (e.g., a study of police over- time assignments last year). Operational analysis coupled with Strategic Crime Analysis, helps patrol commanders to make changes that use resources more efficiently. Used for:

-Assess needs (calls for service, population of data & demographics)

-Generate projections for deployment & resource allocation

**Intelligence Analysis:**

The study of criminal organizations and enterprises, how they are linked, who the key players are. Helps investigation and prosecution units within police. The purpose of intelligence analysis is to assist sworn personnel in the identification of networks and apprehension of individuals to subsequently prevent criminal activity. Used for:

-Linkage between crime organizations & enterprises

-Relate elements such as companies, agencies, people, times, days, to crimes & places

**Investigative Analysis:**

Looks at crime scene, psychological, and forensic analysis used in major crimes. It also helps catch serial killers, arsonists, and similar criminals. The primary purpose of criminal investigative analysis is to develop patterns of serial crimes crossing city, state, and even national boundaries by linking behavior and evidence within and among incidents in order to catch the offender and/or clear cases. Used for:

-Crime scene, psychological & forensic information.

-Link serial or related events.

**SOFTWARE REQUIREMENTS SPECIFICATIONS**

1. **External Interface Requirements**
   1. User Interfaces

The user can interact with the system through a website. The end user can view the visualization of the previous data, based on district and on the type of crime committed. The prediction can also be done extrapolating the previous data and is represented graphically.

* 1. Hardware Interfaces:

Any kind of hand held device or computer with a processor and browser.

* 1. Software Interfaces

R toolkit is used to do the back end processing.

* 1. Communication Interfaces

The communication between the different parts of the system is important since they depend on each other. Internet connection and a web browser are required in order to access the data from the cloud and analyze it.

1. **Functional Requirements**
   1. Functional Requirement 1.1:

ID: FR1

TITLE: View the website

DESC: A user should be able to view the user interface through a browser. The platform should be accessable to all.

DEP: None

* 1. Functional Requirement 1.2:

ID: FR2

TITLE: Select attributes

DESC: The user should be able to select the appropriate attributes like the district, type of crime and other attributes like gender, GDP, population etc.

DEP: FR1

* 1. Functional Requirement 1.3:

ID: FR3

TITLE: Preprocess data.

DESC: Initially a preprocessing step performs a cleaning process in two steps.

It remove data records that are not important for analysis and then implements a missing handling procedure to fill in missing data items or records in the crime dataset.

DEP: NONE

* 1. Functional Requirement 1.4:

ID: FR4

TITLE:Processing of data.

DESC: Clustering of data, classification and deviation detection is done while processing the data.

DEP: FR3

* 1. Functional Requirement 1.5:

ID: FR5

TITLE: Display the results

DESC: Based on the user request, the processed data is visualized as tabulations and graphical representations.

DEP: FR2, FR4.

1. **Software System Attributes**
   1. Reliability

The reliability of the system is the accuracy with which the data is represented. The cloud platform used should be reliable, secure and fast. There should be proper internet connectivity to cloud.

* 1. Availability

The system should be available all the time and care should be taken that the server does not go down.

* 1. Security

The system should be secure enough to prevent the external attacker from entering malicious data which may affect the normal functioning of the software and generate unnecessary exception conditions.

* 1. Portability

The system is portable across different operating systems and different browsers.

* 1. Maintainability

The code written is modular, simple and easy to read. Hence the maintenance required is less. The application need not be updated frequently.

* 1. Performance

The performance of software should be optimized to ensure that the resource utilization is less and efficiency is very high and we get proper results almost instantly with no delays.

1. **Performance Requirements**

The performance of the system depends upon the reliability and efficiency of the servers. It depends upon cloud platform we choose i.e higher the capacity of the cloud, higher the performance. Also a high speed internet is required so as to view the visualization and prediction without much delay.

1. **Database Requirement**

The database should be able to store large number of readings. The database should not crash when the number of hits is high. The database should be flexible and querying should be fast.

1. **Other Requirements**

Knowledge of R, python programming for the processing of data and some web development basics for the user interface.

**DESIGN**

**1. Introduction:**

**A. Number of modules:**

The data mining techniques used in the present work are

 Preprocessing

 Clustering

 Classification

 Deviation detection

**B. Module description:**

Preprocessing:

Initially a preprocessing step that performs a cleaning process in two steps.

1. Remove data records that are not important for analysis.

2. Implements a missing handling procedure to fill in missing data items or records in the crime dataset. [2]

Clustering techniques group data items into classes with similar characteristics to maximize or minimize intraclass similarity. For example, to identify suspects who conduct crimes in similar ways or distinguish among groups belonging to different gangs.

These techniques do not have a set of predefined classes for assigning items. One way is to use the statistics-based concept space algorithm to automatically associate different objects such as persons, organizations, and vehicles in crime records [3] . Another way is using link analysis techniques to identify similar transactions to support the detection and analysis of money laundering and other financial crimes. Clustering crime

incidents can automate a major part of crime analysis but is limited by the high computational intensity typically required.

Classification finds common properties among different crime entities and organizes them into predefined classes.[4] This technique has been used to identify the source of e-mail spamming based on the sender’s linguistic patterns and structural features [5]. Often used to predict crime trends, classification can reduce the time required to identify crime entities. However, the technique requires a predefined classification scheme.

Classification also requires reasonably complete training and testing data because a high degree of missing data would limit prediction accuracy. [6][7]

Deviation detection uses specific measures to study data that differs markedly from the rest of the data. Also called outlier detection, this technique can be applied to recognize fraud detection, network intrusion detection, and other crime analyses. However, such activities can sometimes appear to be normal, making it difficult to identify outliers.

Normally, clustering of crime has a special meaning and refers to geographical grouping of crime. In the present work, it is used to group different kind of crime patters. The result of clustering is then used by the classification process to predict crime trend. Further, it is also process to further identify anomalies from the crime trends. Thus cluster-based prediction crime analyzing tool is proposed in the present research work.

**C. Algorithm design:**

Data Preprocessing:

Most of the data collection techniques like survey studies, field experiments, Crime findings, etc., produce huge amount of information, where missing values are inevitable.

A data preprocessing is a process that consists of data cleaning, data integration and data transformation which is processed by a computer program. It intends to reduce some noises, incomplete and inconsistent data. The results from preprocessing step can be later proceeding by data mining algorithm. The dataset used in experiment contains various items like district of Karnataka where the crime was committed, status, name of the concerning police station, number of crimes with respect to murder, dacoity, riots, rape, burglary etc. , population of the region, GDP, sex ratio and such other demographic factors.

Missing value handling for state field:

The experiment concentrate on only those attributes that are related to crime data, that is year, state, administrative name, number of crimes for the year 2010-2012. The quality of the results of the mining process is directly proportional to the quality of the preprocessed data. Careful scrutiny revealed that the dataset have missing data in state and number of crimes attributes. There are a number of methods for treating records that contain missing values.

1. Omit the incorrect fields(s)

2. Omit the entire record that contains the incorrect field(s).

3. Automatically enter / correct the data with default values (e.g.) select the mean from the range.

4. Derive a model to enter/correct the data.

5. Replace all values with a global constant.

6. Use imputation method to predict missing values.

Missing value handling for number of crimes occurred attribute:

In the present work, while considering the state attribute a method that uses a careful matching technique is proposed.

While considering filling missing number of crimes related murder, dacoity, riots and arson, two methods were used. Initially, all the four fields are analyzed for empty values.

If all the four attributes have empty values for a particular record, then the entire record is considered as irrelevant information and is deleted. While taking individual attributes into consideration, a novel KNN-based imputation method is proposed. In this method, the missing values of an instance are imputed by considering a given number of instances that are most similar to the instance of interest.

The similarity of two instances is determined using a distance function. The algorithm is

1. Divide the data set D into two parts. Let Dm be the set containing the instances in which at least one of the features is missing. The remaining instances will complete feature information form a set called Dc.

2. For each vector x in Dm:

a) Divide the instance vector into observed and missing parts as x = [xo, xm].

b) Calculate the distance between the xo and all the instance vectors from the set Dc. Use only those features in the instance vectors from the complete set Dc, which are observed in the vector x.

c) Use the K closest instances vectors (K-nearest neighbors) and perform a majority voting estimate of the missing values for categorical attributes. For continuous attributes replace the missing value using the mean value of the attribute in the k-nearest neighborhood.

The advantages of using KNN imputation are:

1. k-nearest neighbor can predict both qualitative attributes (the most frequent value among the k nearest neighbors) and quantitative attributes (the mean among the k nearest neighbors).

2. It does not require creating a predictive model for each attribute with missing data. Actually, the k-nearest neighbor algorithm does not create explicit models.

3. It can easily treat instances with multiple missing values.

4. It takes in consideration the correlation structure of the data.

Identification of Crime Zones Using Clustering Techniques:

Given a set of objects, clustering is the process of class discovery, where the objects are grouped into clusters and the classes are unknown beforehand. Two clustering techniques, K-means and DBScan (Density-Based Spatial Clustering Application with Noise) algorithm are considered for this purpose. The algorithm for k-means is given below:

{\underset {\mathbf {S} }{\operatorname {arg\,min} }}\sum _{i=1}^{k}\sum _{\mathbf {x} \in S_{i}}\left\|\mathbf {x} -{\boldsymbol {\mu }}_{i}\right\|^{2}

Where V is the variance, Si is a cluster, ui is its mean, D is the dataset of all points xj.

Partition the dataset into k clusters such that intracluster variance is minimized.

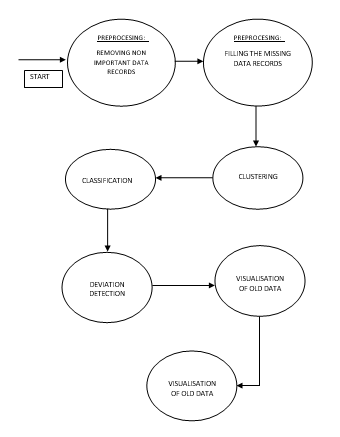
Time and Space Complexity

The primary part of the algorithm is the standard K-means algorithm. Lets us assume that current partition of the N p-dimensional point into k clusters, Compute the distances from each and every point to every cluster centroid and reassign. So for the simplest case of squared Euclidean distance, at every iteration, there is k computations of centroids, each one gets involved in p arithmetic means.

1. **Architecture Design**

****

1. **Flow Diagram**



**IMPLEMENTATION**

**1. TOOL DESCRIPTION**

**R**

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. R is an implementation of the [S programming language](https://en.wikipedia.org/wiki/S_(programming_language)) combined with [lexical scoping](https://en.wikipedia.org/wiki/Lexical_scoping) semantics inspired by [Scheme](https://en.wikipedia.org/wiki/Scheme_(programming_language)). [S](https://en.wikipedia.org/wiki/S_(programming_language)) was created by [John Chambers](https://en.wikipedia.org/wiki/John_Chambers_(programmer)) while at [Bell Labs](https://en.wikipedia.org/wiki/Bell_Laboratories). There are some important differences, but much of the code written for S runs unaltered.Worldwide, millions of statisticians and data scientists use R to solve their most challenging problems in fields ranging from computational biology to quantitative marketing. R has become the most popular language for data science and an essential tool for Finance and analytics-driven companies such as Google, Facebook, and LinkedIn. R includes virtually every data manipulation, statistical model, and chart that the modern data scientist could ever need. You can easily find, download and use cutting-edge community-reviewed methods in statistics and predictive modeling from leading researchers in data science, free of charge. The following features of R make it a powerful tool

**Create beautiful and unique data visualizations**

Representing complex data with charts and graphs is an essential part of the data analysis process, and R goes far beyond the traditional bar chart and line plot. Heavily influenced by thought leaders in data visualization like Bill Cleveland and Edward Tufte, R makes it easy to draw meaning from multidimensional data with multi-panel charts, 3-D surfaces and more. The custom charting capabilities of R are featured in many of the stunning infographics seen in the New York Times, The Economist, and the FlowingData blog.

**Get better results faster**

Instead of using point-and-click menus or inflexible "black-box" procedures, R is a programming language designed expressly for data analysis. Intermediate level R programmers create data analyses faster than users of legacy statistical software, with the flexibility to mix-and-match models for the best results. And R scripts are easily automated, promoting both reproducible research and production deployments.

**Draw on the talents of data scientists worldwide**

As a thriving open-source project, R is supported by a community of more than 2 million users and thousands of developers worldwide. Whether you're using R to optimize portfolios, analyze genomic sequences, or to predict component failure times, experts in every domain have made resources, applications and code available for free, online.

**SAP Lumira**

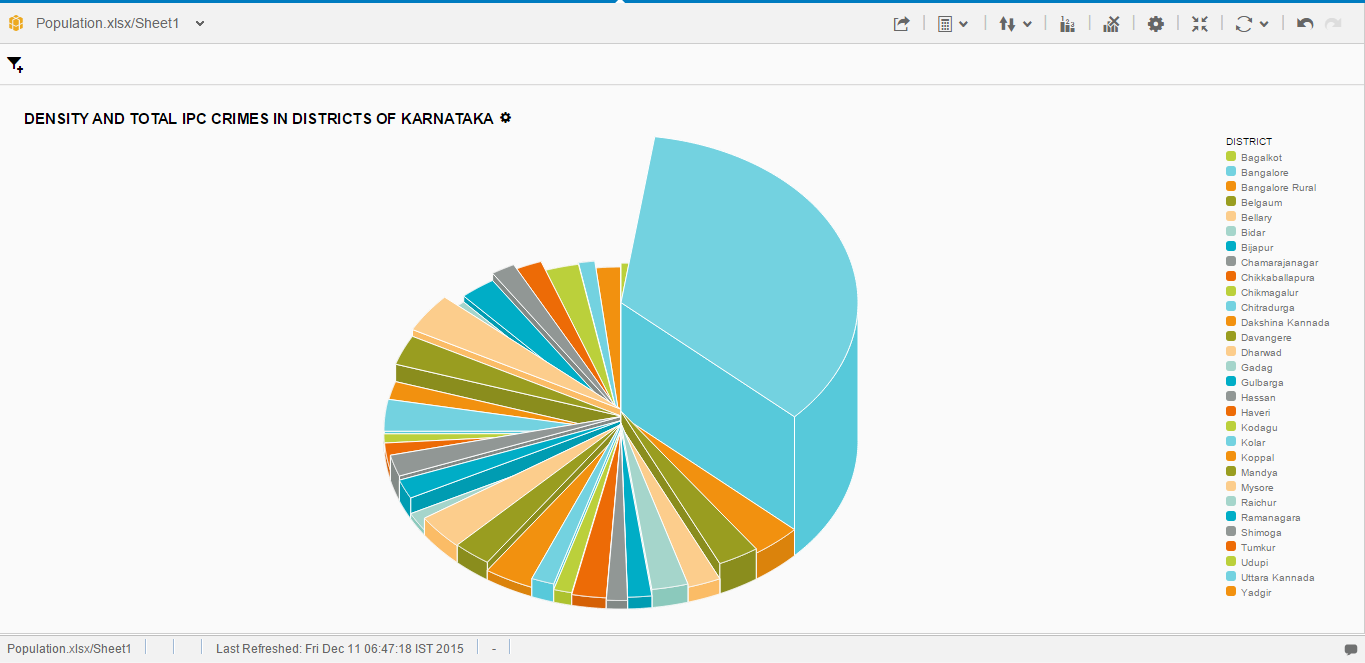
SAP Lumira (formerly called SAP Visual Intelligence) is a self-service, [data visualization](http://searchbusinessanalytics.techtarget.com/definition/data-visualization) application for business users. SAP Lumira grew out of [SAP Business Explorer](http://searchsap.techtarget.com/definition/BEx), the company's previous, but more basic, self-service data visualization tool. According to SAP, the two are  meant to go hand-in-hand. For example, analysts using SAP Lumira can alter data structures and correlations in whatever way they want, without the help of their IT department, then push the data back into the system so it can then be consumed by more casual users in tools like Business Explorer.

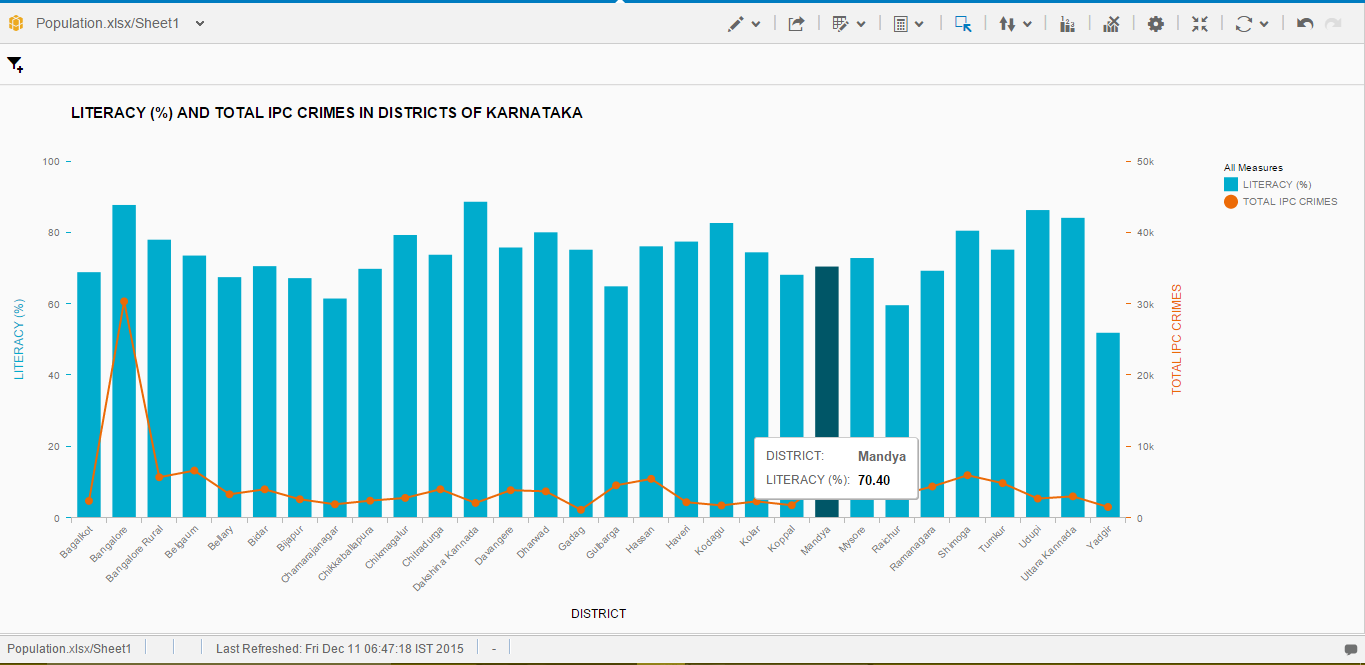
SAP Lumira’s drag-and-drop interface is easy to become familiar with, and with a few clicks data from multiple sources can be gathered, combined, and visualized. After the tool is installed and system access is setup, training is really not needed, but some basics by someone who knows the tool can be helpful.

SAP Lumira’s capabilities include:

* Maximizing business knowledge by integrating wide-scale insights and drill-down to granular details.
* Accelerating decision-making with immediate fact-based solutions to intricate business questions by avoiding list tables and fixed format reports. Interaction is key.
* Increasing self-service data usage without increasing the workload of the IT department
* Visualizing any amount of data in real time using SAP HANA and simple deployment to mobile devices.

**SNAPSHOTS**

****

****

**CONCLUSION**

This project has demonstrated that the linking of crime instance to a location and visualizing the location of crime is possible and the crime analysis can be performed using the crime analysis database. It is possible to extract information about the victims of crime from the database created. It must be pointed out that the data sets in the database such as crime stats in various districts of Karnataka, population, demographics, religion, police stations etc. can be used to locate crime depending on the level of aggregation one is interested in. the database can accommodate such analysis. A number of other information requirements such as when does crime occur (day/night), which crime is common in which region can be provided by using this crime analysis database.

**REFERENCES**

1.Abraham, T. and de Vel, O. Investigativeprofiling with computer forensic log data and association rules, Proceedings of the IEEE International Conference on Data Mining -ICDM'02, (2002) pp. 11-18.

2.Amarnathan, L.C. Technological Advancement: Implications for Crime, The Indian Police Journal, April June. (2003).

3.Brown, D.E. The regional crime analysis program (RECAP): A frame work for mining data to catch criminals," in Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Vol. 3, (1998) 2848-2853.

4.Corcoran J.J., Wilson I.D. and Ware J.A. Predicting the geo-temporal variations of crime and disorder, International Journal of Forecasting, Vol. 19 (2003) 623-634.

5.David, G. Globalization and International Security: Have the Rules of the Game Changed?, Annual meeting of the International Studies Association, California, USA, (2006) http://www.allacademic.com/meta/p98627\_index.html (July 10, 2011).

6.De Bruin, J.S., Cocx, T.K., Kosters, W.A., Laros, J. and Kok, J.N. Data mining approaches to criminal career analysis,” in Proceedings of the Sixth International Conference on Data Mining -ICDM’06, (2006) pp. 171-177.

7.De Vel, O., Anderson, A., Corney, M. and Mohay, G. Mining E-Mail Content for Author Identification Forensics, ACM SIGMOD, Vol. 30, No. 4 (2001) 55-64.

8.Hauck, R.V., Atabakhsh, H., Ongvasith, P., Gupta, H. and Chen, H. Using Coplink to Analyze Criminal-Justice Data, Computer, Vol. 35, No. 3, (2002) 30-37.

9.Keyvanpour, M.R., Javideh, M. and Ebrahimi, M.R. Detecting and investigating crime by means of data mining: a general crime matching framework, Procedia Computer Science, World Conference on Information Technology, Elsvier B.V., Vol. 3 (2010) 872-830.

10.Krishnamorthy, S. Preparing the Indian Police for 21st Century, Puliani and Puliani, Bangalore, India (2003).

11.Marshall, G.M. and Marshall, D.R. Crime in India, Annual Series, 1954-2006, Published by the Government of India, Ministry of Home Affairs, National Crime Records Bureau , Electronic Dataset and Codebook, Published by Center for Systemic Peace. (2008) http://www.systemicpeace.org/inscr/inscr.htm (July 1, 2011).

12.Michelson, M. and Knoblock, C.A. Phoebus: A System for Extracting and Integrating Data from Unstructured and Ungrammatical Sources, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference - AAAI (2006).

13.Nath, S. Crime data mining, Advances and innovations in systems, K. Elleithy (ed.), Computing Sciences and Software Engineering (2007) 405-409.

14.Ozkan, K. Managing data mining at digital crime investigation, Forensic Science International, Vol. 146 (2004) S37- S38.

15.Senator, T.E., Goldberg, H.G., Wooton, J., Cottini, M.A., Khan, A.F.U., Klinger, C.D., Llamas, W.M., Marrone, M.P. and Wong, R.W.H. The FinCEN Artificial Intelligence System: Identifying Potential Money Laundering from Reports of Large Cash Transactions, AI Magazine, Vol.16, No. 4 (1995) 21-39.

16.Tuchinda, R., Szekely P. and Knoblock, C.A. Building Data Integration Queries by Demonstration, Proceedings of the 12th International Conference on Intelligent User Interfaces (2007).

17.Machine Learning & Computational Biology Research Group (2011), http://agbs.kyb.tuebingen.mpg.de/wikis/bg/2.pdf (July 1, 2011).

18.Computer Society of India (2011), http://www.csi-sigegov.org/1/40\_410.pdf (July 1, 2011).