Deviant Behavior Analysis of Juveniles of Detention Centre using Data Mining

A Dissertation Submitted

By

Garima Sharma

(11300009)

To

Department of Science & Technology

In partial fulfillment of the Requirement for the

Award of the Degree of

Master of Technology in Computer Science

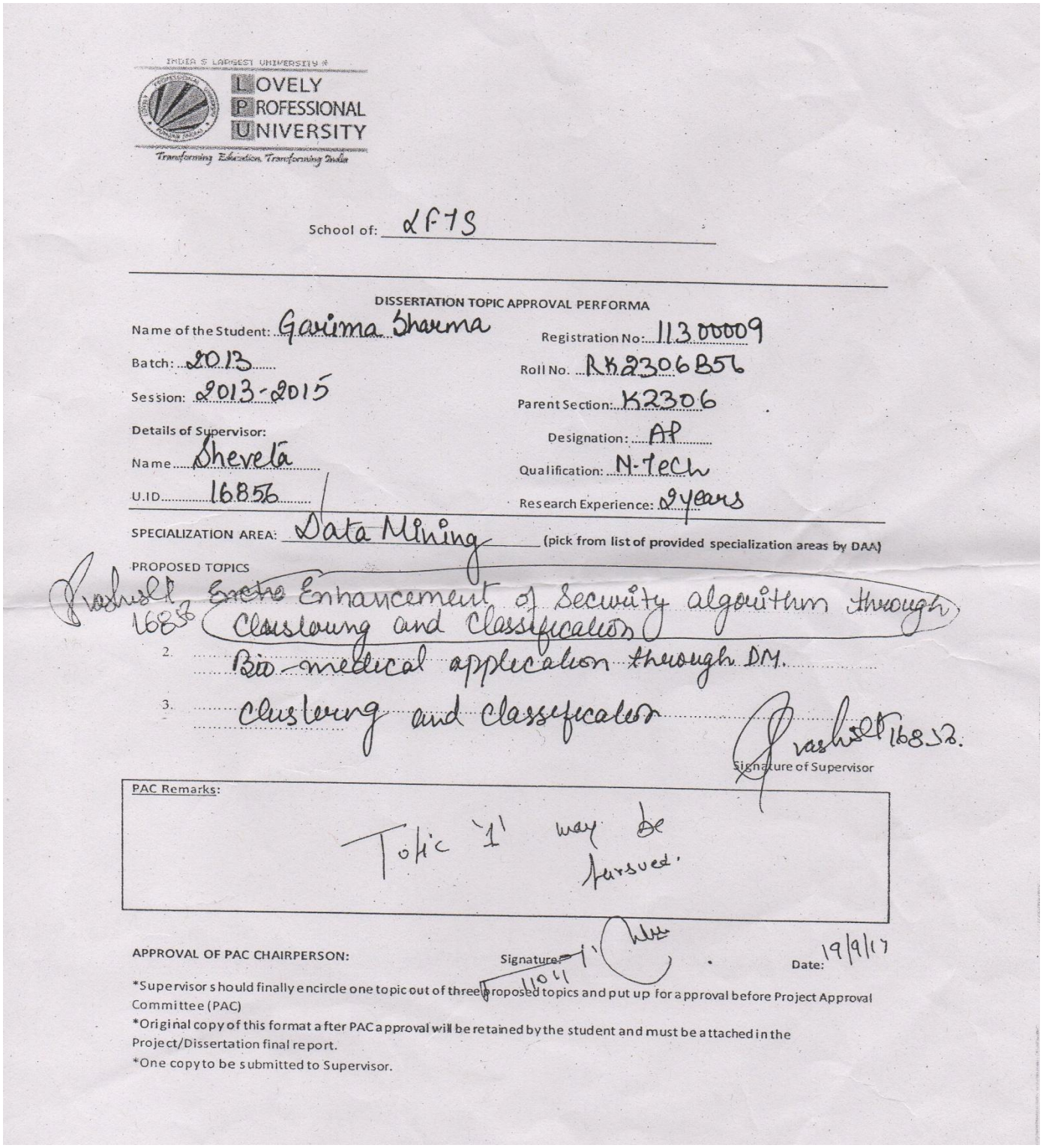
Under the guidance of

Ms Sheveta Vashisht

(16856)

(MAY 2015)

Approval of PAC



INDIA'S LARGEST UNIVERSITY

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

School of: LFTS

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: Garima Sharma Registration No: 11300009

Batch: 2013 Roll No. RK2306B56

Session: 2013-2015 Parent Section: K2306

Details of Supervisor: Designation: AP

Name Shevela Qualification: M-Tech

U.ID 16856 Research Experience: 2 years

SPECIALIZATION AREA: Data Mining (pick from list of provided specialization areas by DAA)

PROPOSED TOPICS

1. Enchancement of Security algorithm through Clustering and Classification
2. Bio-medical application through DM.
3. clustering and classification

Signature of Supervisor

PAC Remarks:

Topic '1' may be pursued.

APPROVAL OF PAC CHAIRPERSON: Signature: Date: 19/9/17

*Supervisor should finally encircle one topic out of three proposed topics and put up for approval before Project Approval Committee (PAC)

*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.

*One copy to be submitted to Supervisor.

ABSTRACT

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data and also known as knowledge discovery from data. Crime is defined as illegal and immoral activity. It is a sin against the society that is often prosecuted and punishable by the law. Crime is the behavior disorder that is an incorporated result of community, cost-effective and environment factors. The development of young people happens in a Hindu joint family. The skill of the students is largely manipulated by education and sociological factors. The growing volume of crime data is the reason of concern about the future of national security. Children are the future of nation. The children who had deviant behavior are now in the correction centers will become the main destructors of nation. The main focus is to analyze the inconsistent and incomplete crime data by using data extraction and data preprocessing to make it efficient and accurate. Data mining clustering algorithm is used for defining the type of crime and classification algorithm used for violence level of juvenile. Since the dawn of civilization the fight against crime has to reduce its intensity in the society. In this thesis work data extraction (DE), data preprocessing (DP) are used to clean and pure database so that data mining techniques can easily applied on them. A clustering algorithm K-means is used for grouping similar level of criminals and classification NEA algorithm is used to predict criminal behavior.

ACKNOWLEDGEMENT

I would like to express my special thanks to GOD to give me this opportunity of writing thesis and providing such nice peoples who was there always to help me in my dissertation. Secondly, big thanks goes to my mentor "**Ms. Sheveta Vashisht**" who gave me this topic "**Deviant Behavior Analysis of Juveniles of Detention Centre using Data Mining**" for my dissertation work, I am heartily thankful to **Sheveta** madam for being my mentor and they also helped me in doing a lot of Research and I came to know about so many new things. At last, I would also like to thank my parents and friends who helped me a lot in my research within the limited time frame.

I would also like to thank my family and friends who have been a source of encouragement and inspiration throughout the duration of this research.

GARIMA SHARMA
(11300009)

DECLARATION

I hereby declare that the dissertation proposal entitled, **"Deviant Behavior Analysis of Juveniles of Detention Centre using Data Mining** " submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: **Garima Sharma**

(11300009)

CERTIFICATE

This is to certified that Garima Sharma has completed M.TecH dissertation proposal titled **"Deviant Behavior Analysis of Juveniles of Detention Centre using Data Mining"** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma. The dissertation proposal is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech Computer Science & Engineering.

Date: **Sheveta Vashisht**

(16856)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

KDD……	**Knowledge Discovery Data**
DM……..	**Data Mining**
DE……	**Data Extraction**
DP……	**Data Preprocessing**
DV……...	**Desired Value**
DS……	**Deviation Standard**
CDCI…….	**Crime Detection and Criminal Identification**
DMT……	**Data Mining Technique**
KDDM……	**Knowledge discovery data mining**
W2T……	**Window Web Technology**
FRBC……	**Fuzzy Rule Based Clustering**
NLP……	**Natural Language Processing**
VSM……	**Vector Space Model**
CA……...	**Corporate Algorithm**
PLA…….	**Principle Component Analysis**
HBV…….	**Hepatitis B Virus**
MV……..	**Missing Values**
SOM……	**Self Organizing Map**
MLRs……..	**Multi-Level Rough Set**
WEKA…….	**Waikato Environment for Knowledge Analysis**

Chapter 1

INTRODUCTION

1. INTRODUCTION

Security is the level of protection against danger, damage, loss, and crime. Security is not new concept in data mining but still there is not too much can be done this area with data mining. The objective of security includes protection of information and property. The growing volume of crime data is the reason of concern about the future of national security Data mining technique can be used in criminal network investigation, security, information sharing system, classification and clustering. Crime is defined as illegal and immoral activity. The growing volume of crime data is the reason of concern about the future of national security. Children are the future of nation. The children who had deviant behavior are now in the correction centers will become the main destructors of nation. The main focus is to analyze the inconsistent and incomplete crime data by using data extraction and data preprocessing to make it efficient and accurate.

With the increasing demand of IT and subsequent growth in this sector, the high- dimensional data came into existence. Data Mining plays an important role in analyzing and extracting the useful information. The key information which is extracted from a huge pool of data is useful for decision makers. Data mining is used for forecasting future trends of market and also helps in decision making like in business, science, medical, counter- terrorism, telecommunication etc. Data mining is also used in analyzing, detection, identification and predicting crime. Crime is defined as illegal and immoral activity. It is a sin against the society that is often prosecuted and punishable by the law. Crime can simply be defined as violate of social norms of particular society. Many situations and circumstances cause crime in the society. From time to time many investigations have been conducted to find out the root cause and way to minimize criminal activities. Fight against the crime is not a phenomenon of contemporary society only rather it has been the prime concern even in the past. Overpopulation, poverty, drugs, family violence, politics, TV violence, depression etc are some causes of crime. Clustering and Classification are

the techniques of data mining are generally used methods of analyzing, detection, identification, prediction of crime.

1.1 DATA MINING

Data mining is the procedure of developing interesting patterns and facts from huge amounts of data and also known as knowledge discovery from data [1]. The data sources can databases, the web, data warehousing etc. Data mining converts raw or huge data into interesting and useful information. The given below steps are involved in extracting knowledge from data:

i. **Data Selection**: The necessary data of required field can be collected from various locations. The data related to the analysis task are retrieving from the database.
ii. **Data Preprocessing**: In this stage, the two processes are involved data cleaning and data integration.
    a) **Data Cleaning**: In this step, noise data, irrelevant and inconsistent data are distant from the collected data.
    b) **Data Integration**: In this step, various data sources of same type are combined.
iii. **Data Transformation**: In this phase, the preferred data is changed into required forms for the mining procedure.
iv. **Data Mining**: It is the important step in which techniques and intelligent methods are applied to extract useful information and data patterns.
v. **Evaluation**: In this step exciting patterns representing facts are recognized.
vi. **Presentation**: This is also an important step, uses visualization and knowledge representing techniques to help users to understand.

The below figure is a outlook from database systems and data mining . Data mining (DM) plays an vital role in the knowledge discovery from Data (KDD). KDD is also known as knowledge pattern extraction, pattern analysis, data archeology, data dredging, information harvest, business intelligence, etc.

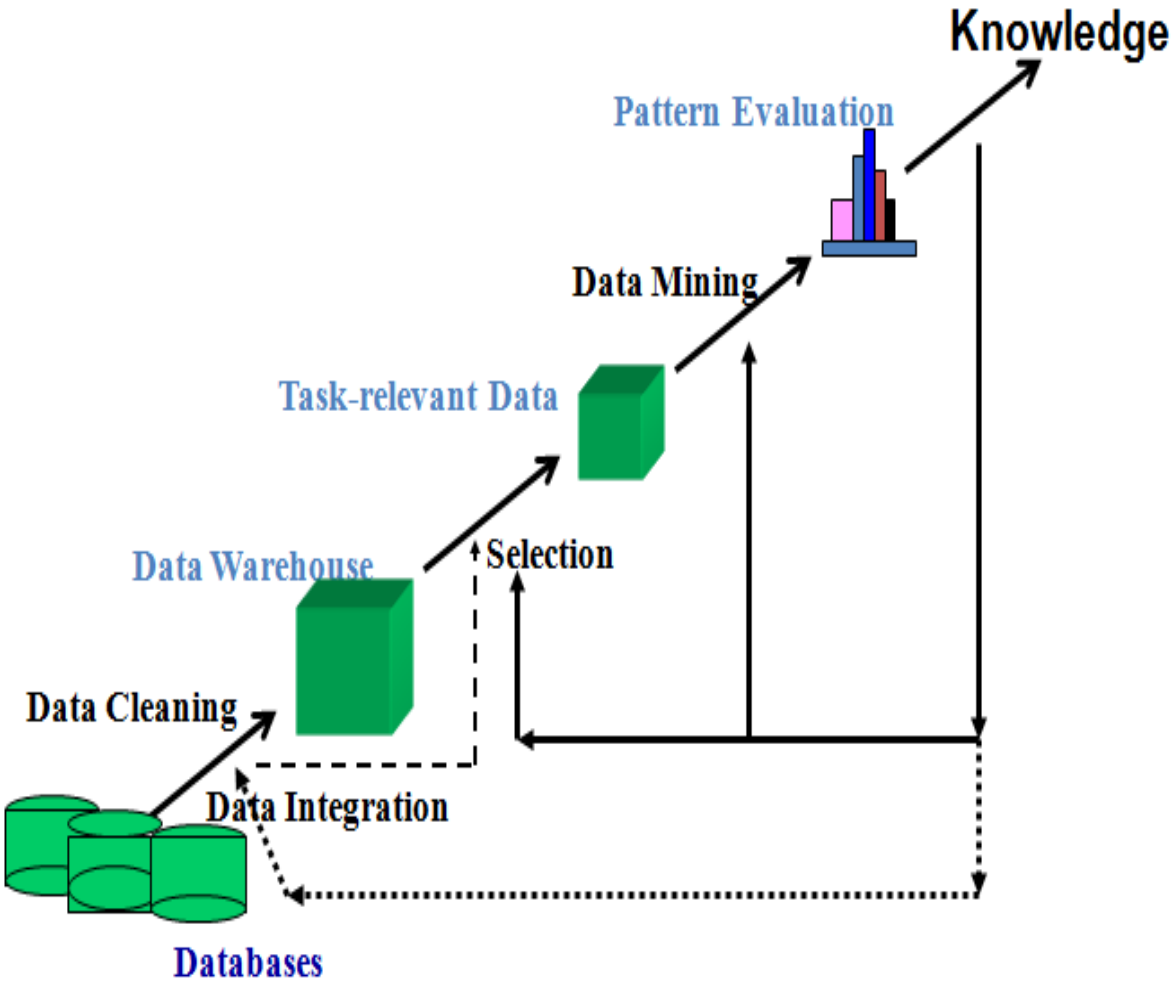


Figure1.1. The process of extracting knowledge from data [1].

1.1.2 DATA MINING TECHNIQUE AND ALGORITHM

Numerous data mining techniques and algorithm like association, classification, clustering, prediction, Regression, Neural Networks etc has been developed and used, Some data mining techniques are discussed below:

1.1.2.1 Classification

Classification is the most commonly used data mining technique, where a model is constructing to predict class. It is used to sort each item in a set of data into one of predefined set of classes or groups. Classification method uses mathematical techniques such as decision trees, linear

programming, neural network and statistics. Classification classifies data items into groups. Types of classification models are Classification by decision tree induction, Bayesian Classification, Neural Networks, Support Vector Machines (SVM), and Classification Based on Associations.

1.1.2.2 Association

Among the various data mining techniques is most commonly and widely used. In association a pattern is formed by using the relationship of a particular item with their item in the same group. To find out the repeated data from large that make us of association and co-relation. The various types of association rules are multilevel association rule, Multidimensional association rule, Quantitative association rule, Direct association rule and Indirect association rule.

1.1.2.3 Clustering

Clustering is a data mining technique that makes useful cluster of objects that have similar characteristic. Classification approach can also be used for distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for selection and classification. Types of clustering methods are Partitioning Methods, Hierarchical Agglomerative (divisive) methods, Density based methods, Grid-based methods and Model-based methods.

1.1.2.4 Prediction

Building relationship between independent and dependent variables is known as Prediction in data mining technique. For prediction we can make use of regression technique where regression analysis is helpful in modeling the relationship between one or more independent and dependent variables. Types of regression methods are Linear Regression, Multivariate Linear Regression, Nonlinear Regression and Multivariate Nonlinear Regression.

1.1.3 EVOLUTION OF DATA MINING

In early 1960s, the data can be collected on Disks, tapes and in Computers which were provided by IBM and CDC and data were Retrospective and Static data. The functionalities were like data collection, creation and management. In 1980s, Oracle, Sybase, IBM, Microsoft provided technologies like relational databases, SQL, Hierarchical and network database systems. The Data were Retrospective and dynamic data at record level. In 1990s, the data were Data Warehousing, Decision Support System, On-line analytic processing (OLAP), multidimensional databases and data warehouses which were provided by Pilot, Comshare, Arbor, Cognos, Microstrategy and data were Retrospective, dynamic data delivery at multiple levels. Today's, advanced algorithms, multiprocessor computers, massive databases are technologies which are provided by Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry) and data are in the form of Prospective and proactive information delivery. The evolution of the data mining started in 1960‟s when the data was recorded on the tapes and disks and slowly with the advancement in technologies the concept of data mining got wider. Now day's massive algorithm, multipurpose algorithms are included in the data mining.

Table 1: Steps in the Evolution of Data Mining.

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases ,SQL	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing &Decision Support	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups	Prospective, proactive information delivery

1.2 JUVENILE CRIME

Definition of crime varies from society to society and place to place. The Government of different countries and different political organization are busy forming out new and effective policies to lower down the crime rate. They are putting maximum efforts to provide safe and peaceful living for the people. It is not only the concert of people in modern era but since the conception of society, it has been the prime concern. Some of the causes of crime are: (a) Gang involvement, (b) Weapons power, (c) Absence and school dropout, (d) Drug violence and immature drinking, (e) Recidivism among youth on trial, (f) Poverty, and (g) Poor parenting. Those children indulge in criminal activities who do not get any intention in the family. They involve in crime to win their attention some time violence in families also leads to crime. A person under the state of hopelessness and frustration very easily resort to violence and thus harms himself and others. A drug addict indulges in crime to raise money and satisfy his need. Mostly children are involved in drugs trading also. The declassification of children in correction centers (juvenile prison) is also a major cause of crime creation. There are different types of factors involved in increasing violence in juveniles like Individual factor, School factors, Society factors etc.

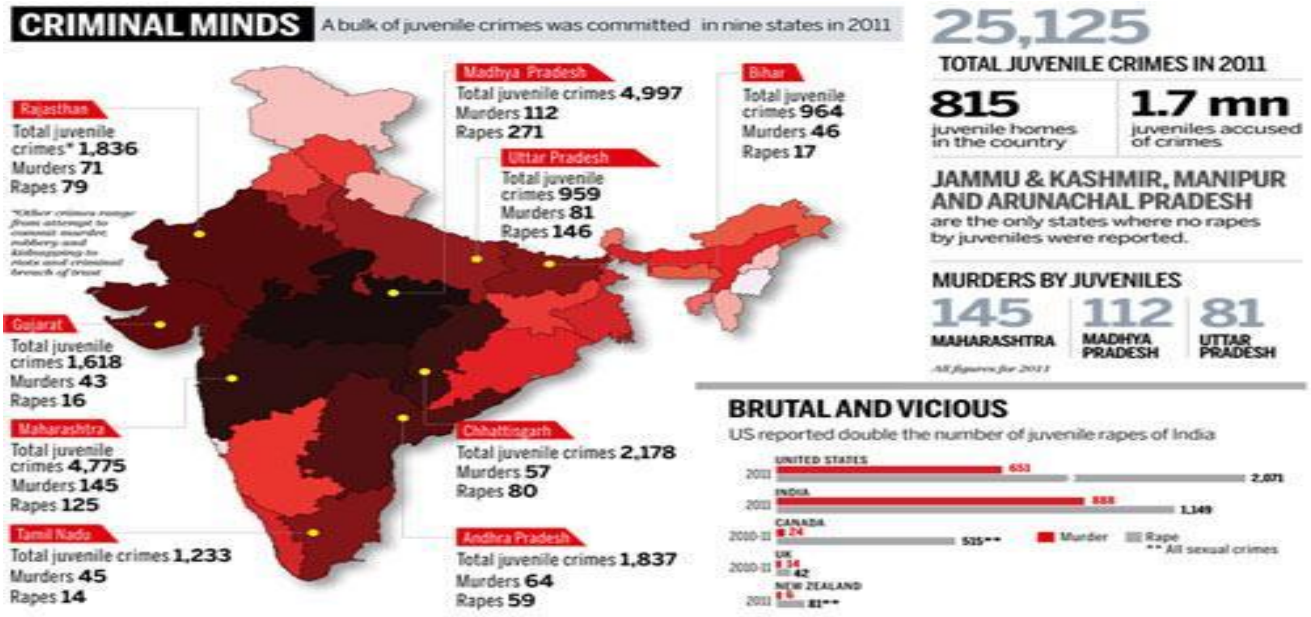


Figure 1.2: Juvenile Crime committed in 9 states in India [9].

As the country have to to revise its laws on juvenile age because day by day the ratio of criminal activities are increased very fastly by juveniles as per the statics of the National Crime Record Bureau.

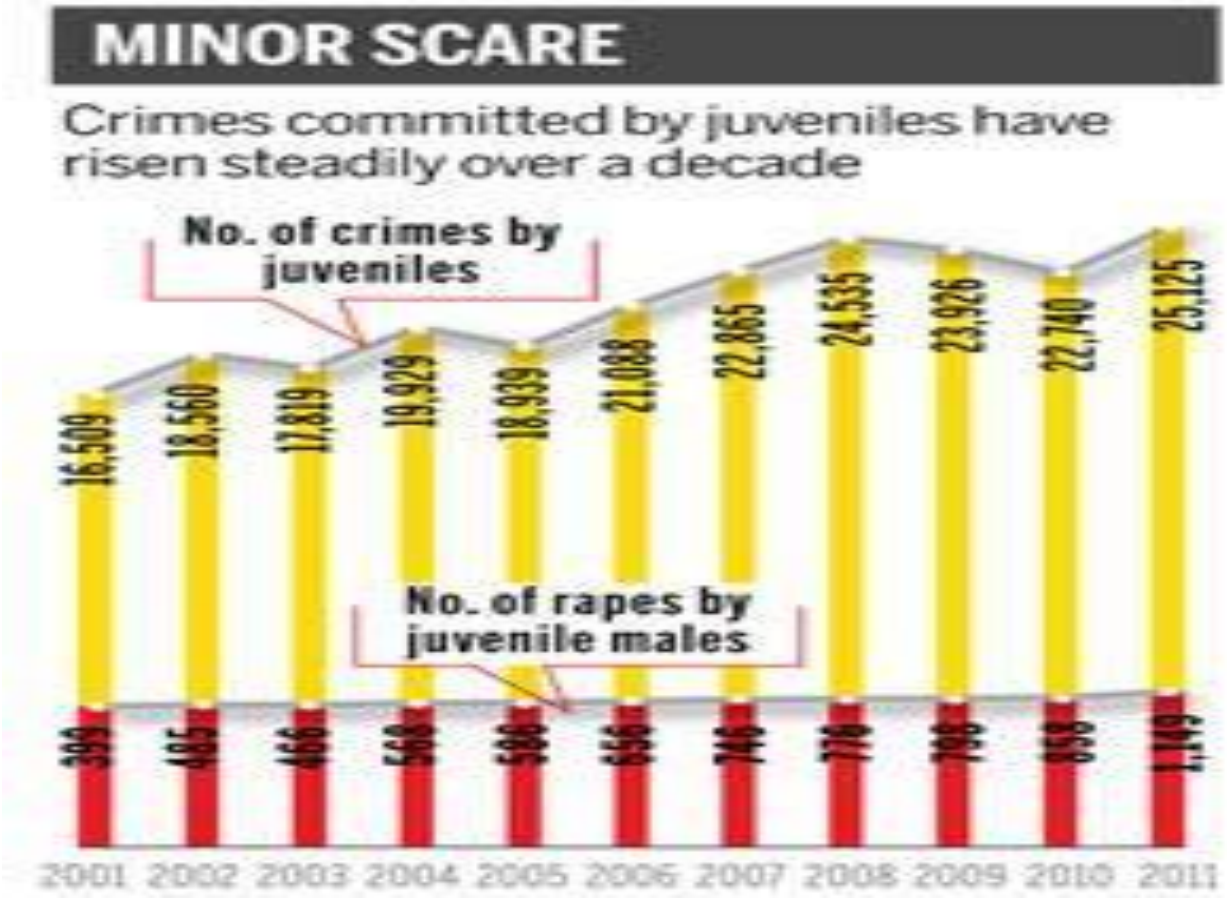


Figure 1.3: Raising Ratio of Criminal Activities by Juveniles [8].

As the country have to revise its laws on juvenile age, Punjab's crime graph states that the quite young are committing shocking crimes such as rape and murder. [4] 64% of juveniles in custody for various crimes fall in 16 to 18 years age. 7 out of 10 rapes committed by juveniles in Punjab last year were by 12 to 16 year olds.10 juveniles were in custody last year on murder charges, of which one was in the age group seven to 12, three between 12 and 16 and six in the 16 to 18 group -- those charged with rape, kidnapping and higher.

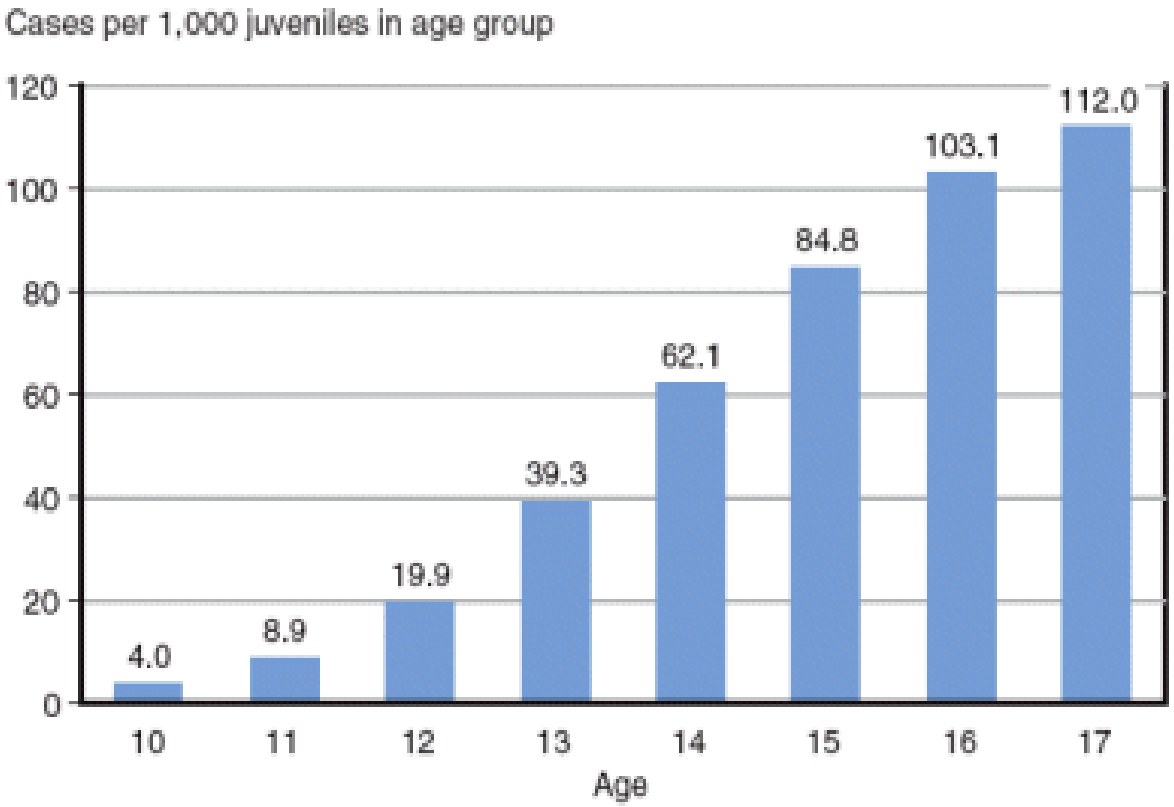


Figure 1.4: Statics of Juveniles Age Groups [5].

1.3 Detention center

Detention centre is a type of Borstal school. It is an body which functions primarily under the provisions of the Borstal Act, where juvenile prisoners, from the age of 08 years to 18 years are kept. It is used totally for the custody of immature. The mainly purpose of Detention centre is to provide concern, safety and psychotherapy of juveniles. Detention center is mainly used to provide an environment suitable for children development that are involved in criminal activities and keep away from environment of the prison. The Borstal Schools provided a variety of professional guidance and tutoring with the help of qualified teachers. The main motive is give culture, guidance and improve morally so that juvenile can restructure their life and prevent themselves from crime.

In India there are ten States namely, Andhra Pradesh, Haryana, Himachal Pradesh, Jharkhand, Karnataka, Kerala, Maharashtra, Punjab, Rajasthan and Tamil Nadu having Detention centers. Tamil Nadu detention centre has the highest capacity of keeping juveniles. Haryana, Himachal Pradesh and Punjab have separate female detention centers.

In Punjab, one borstal school in Ludhiana city and one female detention center in Jalandhar known as Nari Niketan.

Chapter 2

REVIEW OF LITRETATURE

Shady Shehata *et al* (2010) in this paper Text mining is analysis on the basis of statistical. A concept based model is use to analyze the sentence and document because statistical analysis only capture the frequency with in the document. The concept based model can easily detect the unimportant factor in the document. The proposed model can find the similar frequency text. The model is consisting of sentence, document, corpus and concept based measures. In text clustering the large set of experiment can do. The four components sentence, document, corpus and concept is involved to enhance the text clustering [15]. The sentence based model is used to analysis the semantic structure of the sentence to measure frequency. The document based model is used to analysis the concept of document. The corpus based model is used to analysis the concept of document frequency. The concept based model is used to analysis the concept as well as the semantic of the sentence and document. Clustering, one of the traditional data mining techniques is to identify inbuilt groupings of the text documents. Most current document clustering methods are based on the Vector Space Model (VSM) as well as the natural language processing (NLP) which is used data representation for text classification and clustering. The efficiency and the accuracy of text clustering can be achieved by the proposed model.

Ji Dan *et al* (2010) in this paper an incorporated algorithm of data mining named as CA was developed. This algorithm improves the initial methods of C4.5 and LURE. The algorithm uses the principle component analysis (PLA), parallel processing and grid partition to get reduction of feature and scale for data sets which are large. The researchers have developed this algorithm for maize seed breeding and their experimental result proves that original methods are not that better as their approach. For this research they have assembled large amount of agricultural information data which is used for vast territory and diversity of crop resources. Due to agricultural distinctiveness such as crop resources complexity, consequences among thickness, climate, fertilize thickness and lack of useful tools researches used only small quantity of data. The main objective of their research is to help people in order to analyse and collect useful information for seed breeding .And according to them data mining development to agriculture is a new research point.

Kwong-Sak Leung *et al* (2011) in this paper meaningful information from large datasets. The major challenge is to identify the Hepatitis B Virus (HBV) that is the type of HCC (liver cancer). The data mining techniques is used to detect liver cancer. The two methods are used rule learning and non linear integral classification methods are used. The fuzzy measure is used to find 70% accuracy and 80% of sensitivity of datasets [16]. Clustering methods are used for separating the subgroups Evolutionary algorithm is used. Classification methods used in this paper are non-integral and fuzzy measures. The datasets are splitting into two parts training examples and testing examples. In this paper classification totally on non linear integral to avoid overtraining.

Eghbal G. Mansoori (2011) in this paper Fuzzy clustering is better than crisp clustering when the boundaries among the clusters are unclear and confusing. To overcome the limitation of both the clustering a novel fuzzy rule-based clustering algorithm (FRBC) is proposed [17]. The fuzzy and crisp clustering is sensitive to potential clustering. The fuzzy rule-based classifiers are a supervised classification technique to do unsupervised clusters analysis. The fuzzy rule based algorithm is used to represent the knowledge and data of clusters even potential clusters. The traditional algorithm only determines the center and member of clusters. The fuzzy rule represents the knowledge and data both in understandable form. The novel fuzzy rule based algorithm gives more accurate and efficient result.

Malathi. A *et al* (2011) in this paper, the clustering algorithm is used for predicting crime pattern from missing values and fast up the process of solving crime. For filling missing valves and predicting of crime pattern, MV algorithm and Apriori algorithm is used with some enhancements. The Data mining techniques are applied in the field of criminology to extract useful information from huge volume of data of criminals. The information like criminal behavior and relationships between crime patterns. The proposed model of prediction crime pattern is very useful for Indian scenario to handle crime investigation efficiently. The proposed tool is very efficient in handling large datasets of crime data. In this paper, mainly four steps are involved data cleaning, clustering, classification and outlier detection. [18]Experimentally proved in this paper the proposed tool is effective for Indian police for speedily analysis of data, identify crime pattern and predict crime.

Lijun Wang *et al* (2012) in this paper ECKF, a proposed framework for evolutionary clustering [19]. ECKF has computational efficiency and hence is applicable to large evolutionary datasets. Mathematically proved the union and accuracy of ECKF, and provide detailed analysis of its computational efficiency (both time and space). In a distinctive large data mining application, data is not only collected over a period of time, but the connection between data points can change over that period too. Moreover, this change may either be extreme (e.g., a large of number of active users adding each other) or slow (e.g., a small set of users slowly become inactive). An occurrence of this kind typically is due to the presence of noise, and the algorithm should be strong enough to overcome it. This dual objective evolving nature of clustering is very different from the goal of a traditional clustering algorithm, and its falls in the paradigm of evolutionary clustering. The size of the data poses it's also a challenges. It is difficult and in many cases even computationally infeasible to re-cluster as the large-scale data evolves.

Wei Wei *et al* (2012) in this paper the genuine customer and fraud user detection from highly imbalanced data. Window Web Technology (W2T) methodology is used to detect fraud. In financial crime management, the online banking fraud is becoming a major issue. Traditionally the detection system investigation handles manually which is very time consuming. Three models are used Cost sensitive neural network, Contrast pattern mining, and decision forest. [22]There are many fraud like fraud in online banking, credit card fraud etc. three steps are involved in this framework are database, preprocessing and modeling. Contrast pattern mining used for analysis behavior, cost sensitive neural network used for comparing fraud and genuine transactions, and decision forest used for making decision.

Jasna Soldic-Aleksic *et al* (2012) this paper have provided the results of the application which is a combination of two models of data mining namely Kohonen Self- Organizing map (SOM) and CHAID. The result provided was for the problem of clustering in the marketing sphere. Kohonen SOM model was used by the researcher for visualizing and making clusters of market data. Further the researcher used the results for the analysis using the CHAID algorithm. The combination of two models was used because according to the researcher CHAID model is an efficient interpreter of visuals for the cluster results depicted by SOM. This two phase technique can be used in studying various aspects of markets, as open survey of

market will help to get the output according to customer needs because customer needs can be studied easily by using this approach of two different models which are combined together.

Gunjan Mansingh *et al* (2013) in this paper, organizations can enhance decision making system by analyzing the datasets through data mining techniques. Knowledge discovery and data mining (KDDM) process is used to process the datasets and can be easily understandable by users. A set of model can used to take effective decision in every phase. The phases are business problem understanding, understanding and capturing the data, applying data mining techniques, and interpreting results [23]. The main thing in this paper is the study of both the demographic and attitudinal behavior. The KDDM process provides the systematic and structured manner to data so that the results should be more accurate and reliable. Internet banking used for opening new deposit accounts, online transaction, transferring funds such as electronic bill payment and receipt. The objective of this paper is to analysis the level of usage of different internet banking and fining segments of data. The thirteen set of models such as nine decision tree and four logistic models are used analyze the performance measures.

Mingquan Ye *et al* (2013) in this paper an innovative multi-level rough set model (MLRs) which is based on attribute value taxonomies and a program of full sub tree generalization is presented. The researchers have compared the results of MLRs with that of the Pawlak's rough set model. Along with this another different concept of cut reduction which is based on MLRs is introduced. According to researchers a cut reduction has the ability to reduce the multi-level decision table which is more abstract, the reduction is done with the classification ability which is same on the decision table which is raw. The main focus of the researchers is to enhance the simple-level Pawlak's rough data set model on a concept of multi-level rough set model(MLRs).The cut reduction in MLRs evaluation has n-hard problem and for computing the cur reduction a CRTDR algorithm is presented. The experimented results of the research proved the powerfulness of the methods proposed by the researchers. Further they have researched on the extension of the proposed model for discovering multi-level decision rules and how other rough sets can be extended in association with attribute value taxonomies.

Raed T. Aldahdooh *et al* (2013) in this paper describes the method to identify initial clustering centroid of K-mean clustering. Previous approaches use random selection method for

identifying initial centric. This paper enhances the performance of initialization method over many datasets by taking into consideration different observations, number of clusters, groups and clusters complexity. The experimental results shows that the proposed initialization methods lead to better clustering results than random method and are very effective. The paper works on following strategy to find out the initial centric: Firstly, select the initial centroid by using random method. After selecting initial centroid, some calculations are performed to determine the initial centroid and the points which are closest to the centroid. The choice of the calculations performed is based on finding the Euclidian distance or some other points found on the dataset. To find the number of nearest points of the initial centroid divide the total no. of objects used in the given dataset by the total number of the clusters given by the user. If the first selected centroid points contain the noise, then it was ignored. Another point is selected until the first centroid point is not found.

Jyoti Agarwal *et al* (2013) in this paper a crime analysis is done using K-mean clustering using Rapid miner tool. The main objective of doing crime analysis involved in this paper is to predict the crime causes based on the existing crime data and apply data mining techniques in an efficient manner. The tool used in this paper mine the dataset according to user requirement and applies K-mean clustering to compute the distance matrix. After then crime analysis is applied on the resultant cluster. From the clustered result, it is observed that the crime rate decreases over the years and this approach is useful for finding the new precautions method for future

Suwimon Vongsingthong *et al* (2014) in this paper social networking sites especially facebook is very popular among students. These social networking sites are used for making new links and/or connections, sharing unstructured and semi-structured data. In this paper, the student behavior gets analysis about "share" on facebook. What the student mostly like things like ads status etc and share them on their own timelines. For analysis data, six clusters can make like dining, itinerary, pets, entertainment, games, and gifts/varieties. For classification KNN, Decision Tree, Naïve Bayes and SVM are applied to find the mostly liked product [20]. The manufacturing companies should post ads on the social networking sites and get the review of the public especially university students. The taste of the youth can be easily extracted by

analyzing the social networking sites. The experimental research shows university students mostly liked entertainment products like music and movie CDs and then pets.

Devendra Kumar Tayal *et al* (2014) Crime in India is increasing with very fast growing rate. Some factors like poverty, unemployment, frustration, corruption and illiteracy are playing very in increasing rate of crime. In this paper, author presents a framework to contribute toward decreasing crime rate and to identify criminals. They propose a technology of crime detection and criminal identification (CDCI) using data mining techniques (DMT) for Indian cities [21]. They analysis seven cities and collected unstructured crime data from crime based web sources. Use of K-means and KNN algorithm of clustering and classification for extracting and make unstructured data into structured form. In this paper work is divided into four cases: case (1) detects crimes in India, case (2) detects crime specific location, case (3) detects crime of specific type, and case (4) detects crime of specific type and location with the help of Google map through Netbeans. WEKA is used to verify clustering. The results are accuracy of 03.62% and 93.99%, respectively. In future, data privacy, reliability, accuracy and other security measure of crime based data mining system can be enhancing. Data mining can be used to detect criminal problems. Any research that can help in solving crimes faster will pay for itself. About 10% of the criminals commit about 50% of the crimes. The use of clustering algorithm for a data mining approach to help detect the crimes patterns and speed up the process of solving crime.

Gilad Katz *et al* (2014) in this paper a method was developed named confDtree (confidence-based decision tree) which can be used for the three drawbacks of decision tress. According to the researchers there are these problems which effects decision trees which are performance reduction while dealing with the small training set; criteria of decision tree is very solid and exact; and that a single uncharacteristic attribute sometimes results in derailing of the process of classification. ConfDtree is a post processing method which has the liberty to classify the instances outliers of decision tress in a better way. The researcher stated that the predictive performance of decision trees is increasing steadily and powerfully. The average improvement calculated for minor, in equal or multi-level class dataset is from 5% tom 9%.When reported in the performance of AUC. For making the method able to select appropriate algorithm for particular dataset along with maintaining the gained benefits which are introduced by using confidence intervals, it is important that the method has the facility to integrate with every

algorithm of decision tree. There are mainly two drawbacks of the proposed algorithm: firstly the algorithm makes a small increase in the computer that cost used for classification of new instance and secondly the reduction of the comprehensibility of the model is done by also.

Chapter 3

PROPOSED WORK

3.1 SCOPE OF STUDY

This research work focus on rising a tool for Indian scenario using different data mining techniques that can professionally handle crime investigation. The fight against crime is not a new thing in society and it was from the establishment of society. The proposed research tried to bring crimes rate down. The proposed research enables agencies too effortlessly and cost-effectively clean, describe and analyze incomplete and inconsistent crime data to identify type of crime, criminal behavior and violence level of a criminal. The proposed research, applied to crime data, can be used as a knowledge discovery framework that can be used to analysis really large datasets and integrate a infinite group of methods for accurate handling of security issues. Data mining clustering algorithm is used for defining the type of crime and classification algorithm used for violence level of juvenile to decrease the crime rates in the society. The development of young people happens in a Hindu joint family. The skill of the students is largely manipulated by education and sociological factors. The growing volume of crime data is the reason of concern about the future of national security. Children are the future of nation. The children who had deviant behavior are now in the correction centers will become the main destructors of nation. The main focus is to analyze the inconsistent and incomplete crime data by using data extraction and data preprocessing to make it efficient and accurate. Data mining clustering algorithm is used for defining the type of crime and classification algorithm used for finding deviant behavior of juvenile.

3.2 OBJECTIVES

Objective will give importance on the aim of the research work. The Government of different countries and different political organization are busy forming out new and effective policies to lower down the crime rate. They are putting maximum efforts to provide safe and peaceful living for the people. It is not only the concert of people in modern era but since the conception of society, it has been the prime concern The main focus is to analyze the inconsistent and incomplete crime data by using data extraction and data preprocessing to make it efficient and accurate. Data mining clustering algorithm is used for defining the type of crime and classification algorithm used for deviant behavior of juvenile in the correction centers. In this research work data extraction (DE), data preprocessing (DP) are used to clean and pure database so that data mining techniques can easily applied on them. A clustering algorithm is used for grouping similar level of criminals and classification is used to predict deviant behavior of juvenile. The output of the system is helpful in taking decision about the juvenile's future that is in the correction centers.

1. To extract and preprocess the incomplete and inconsistent data.
2. To accurately and efficiently analyze the growing volume of data.
3. To explore and enhance clustering and classification algorithm to identify type of crime.
4. To construct an efficient framework for predicting deviant behavior of juveniles

3.3 METHODOLOGY

3.3.1 PROBLEM FORMULATION

. The development of young people happens in a Hindu joint family. The skill of the students is largely manipulated by education and sociological factors. The growing volume of crime data is the reason of concern about the future of national security. Children are the future of nation. The children who had deviant behavior are now in the correction centers will become the main destructors of nation. . It is not only the concert of people in modern era but since the conception of society, it has been the prime concern. There are four types of factors involved in increasing violence in juveniles.

3.3.1.1 Individual factor

An increasing the risk of violent behavior such as competition, earnings, gender, and family structure are the factors involved. Recently, the national survey study of adolescent health is too weak which explaining the youth risk behaviors. Poor academic performance, unstructured free time is to engage in violent behavior. Criminal peer's relationship with friends who connect in risky behaviors is linked to students' later involvement in violence. The violence of home and whose parent is criminal impact more on the children thinking and they will attract towards crime.

3.3.1.2 School factors

In school, the competitions between the children grow the violence in them. The parents whose are working are so busy and they don't have much time to sit and talk with their children's. In order to gain attention of their parents and friends children always do small mistakes in homes and as well as in schools and grow criminal behavior in child. The violent movies and serial mostly influence the children. What they see in TV can apply them in real life.

3.3.1.3 Society factors

Inferior schools and housing grow feeling between youth that society does not be concerned about them. Violence can then become an face of students irritation and isolation. Youth people think that the violence is the easy and only way to gain attention. The craze among youth of

doing something exciting and new throws them in the way of crime. The use of weapons like guns and tasting of drugs increase the violence in them. The mindset of being rich and famous in the society leads the youth toward the adoption of crime. One study shows that 40% of students of schools drunk alcohol because the status in the society.

3.3.1.4 Other factors

Students may experience unkindness based on their age competition, gender, and the sensitivity that they cannot contribute to society in meaningful ways. Poverty, high unemployment, and lack of necessary resources can create a sense of hopelessness among both youth and adults. Stressful family environments, conflict in the home, lack of fathers in the home, insufficient parenting skills, and poor communication can contribute to students' feelings of unimportance which can give birth to violence. Mental illness and disorders affect students' ability to learn, communicate, and make good decisions. They can harm themselves as well as others.

The prediction of violence level in juvenile can do by understanding juvenile behavior. In this research an enhanced clustering and classification techniques can use for predicting violence level in juveniles.

3.3.2 RESEARCH DESIGN

3.3.2.1 The design for overall process

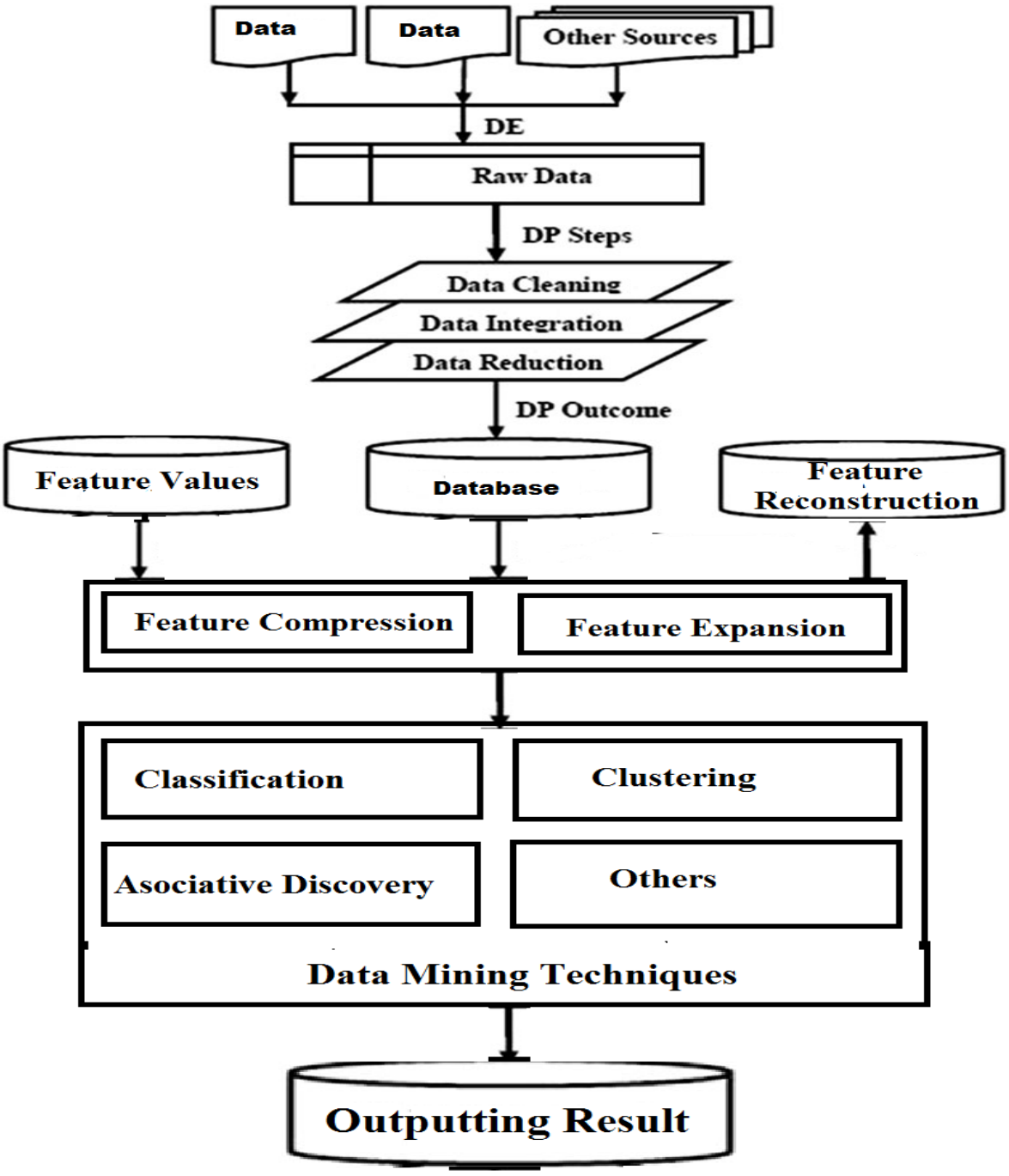


Figure 3.1: The Overall Process

The process in Figure mainly includes the following modules.

- **Collection of data**: The data can be collected from different sources like questionnaires, survey and etc. Collection of data from different sources is known as data extraction. That data is known as raw data.

- **Data Pre-processing**: The processing which is performed on raw data to prepare it for another processing method is described by data processing. Data cleaning and Integration can be done on raw data. Missing values can be filled Data preprocessing transforms the data into a pure data known as database.

- **Obtaining Database**: After applying data pre-processing, the outcome is database that is pure data.

- **Feature reconstruction**: The merging of some feature and expansion to append number of features refers to feature compression.

- **Data mining:** Clustering, classification, retrieving, associative discovery and many more are categorized under data mining. These modules and techniques like data extraction, preprocessing are featured of reconstruction and this data mining is the main focus of study.

3.3.2.2 The design for Classification

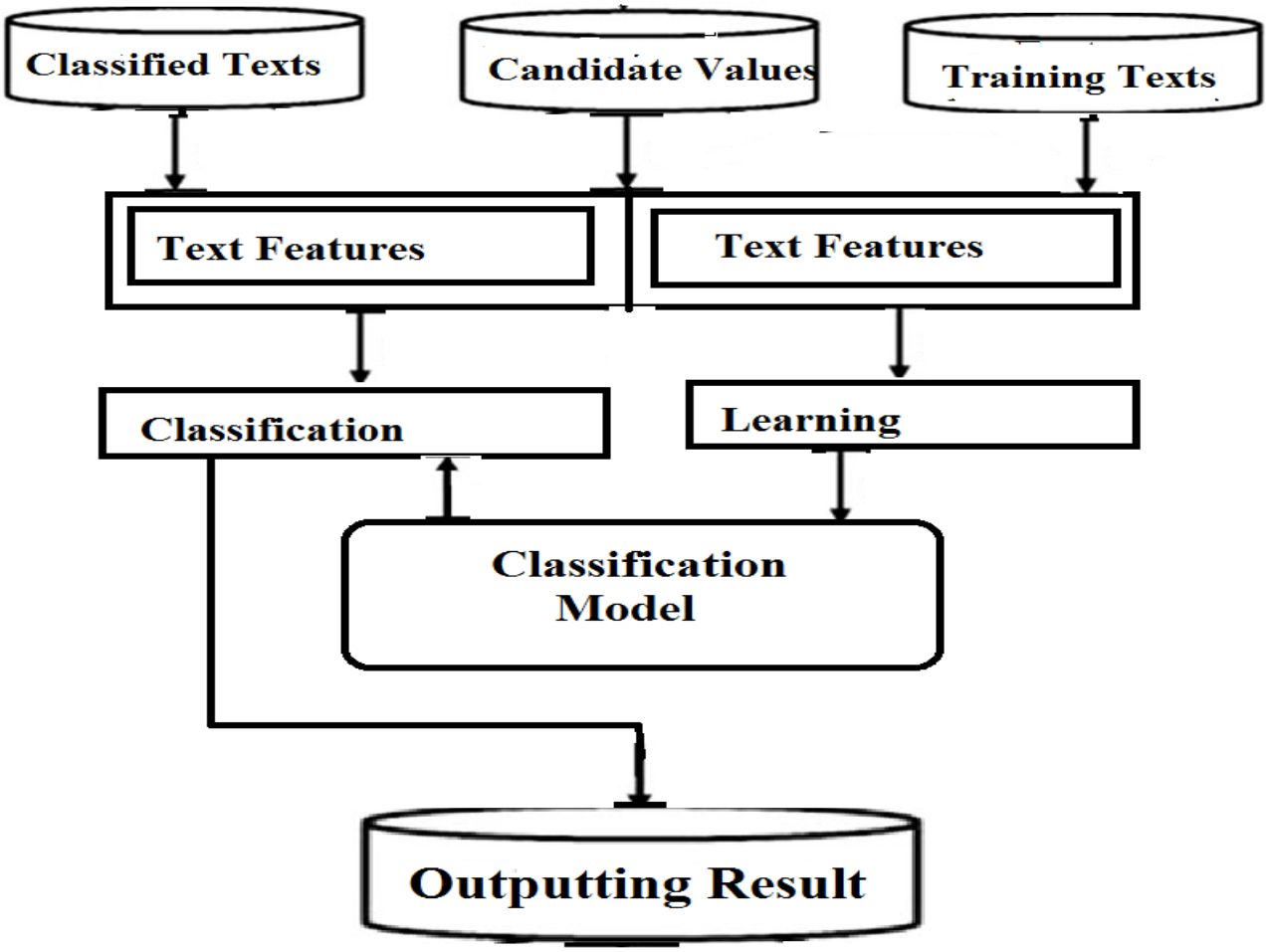


Figure 3.2: The Model of Classification

As shown in Figure 3.2, the model of text classification mainly includes two parts:

- **Learning:** The memorization [10] of patterns and the subsequent response of the network can be categorized into two general paradigms:

(a) Supervised learning [11,2,3] is a machine learning technique used for creating a purpose from training data. The training data consist of sets of input things, and outputs. The output of the function can predict a class label of the input object (called classification).

The task of the supervised learner is to predict the value of the function for any input object after having seen a number of training examples (i.e. pairs of input and target output).

(b) Unsupervised learning [11, 2, and 3] is a method of machine learning where a structure is used to fit to observations. In unsupervised learning, a data set input things is gathered. Unsupervised learning treats input objects as a set of random variables.

Classification: Classification is a data mining (machine learning) technique used to predict group membership for data instances.

3.3.2.3 Attributes

1. Crime_type:

In this, the different type of crimes are present which are typically done by juveniles like murder, attempt to commit murder, theft, Burglary, Arson, rape, Robbery, molestation, sexual harassment, Kidnapping, Riots. Total eleven different crime can taken.

2. Month:

In which month the crime is done from month January means 1 to December means 12.

3. Year:

The data is of two years that is 2011 and 2012.

4. crime_count:

It means in particular month and year, the same crime occurred how many times.

5. Age:

The juvenile criminal age is from 7 years to 18 years but as per the statics mostly crime is done by the age group of 12 years to 18 years.

6. Divisions:

There are total eleven divisions of police station in Jalandhar.

D1- Industrial area

D2- Patel chownk

D3- Milap chownk

D4- Jyoti chownk

D5- Bharkon champ

D6- Model town

D7- Garha road

D8- Focal point

D9- Baradari

D10- Basti bawa khel

D11- Sadar thana

7. District:

The data is collected from eleven divisions of jalandhar district of Punjab.

8. Location:

On which location the crime is done like Home, shop, office, road, school, bus stand, station, train.

3.4 BASIC IDEA

The proposed research makes use of K-means clustering algorithm and NEA (New Enhanced Algorithm) classification algorithm. NEA has two stages. First stage is to cluster the nature of crime and analysis of clustered database after applies NEA classification.

A. Definition and Notations

Let D is the database which contains N number of transactions‟= {d1,d2,d3……di,dj…..dN}, di,dj are two records in D.

K: Number of clusters.

1. Square error E It is the measure similar type of cluster, which is based on the distance between the object and the cluster mean.

2. Entropy It is a measure of the impurity in a collection of training samples.

3. Information Gain It is the impurity degree of the parent table and weighted summation of impurity degrees of the subset tables.

B. K-means Clustering Algorithm

This type of clustering comes under centric models of clustering. This algorithm belongs under the family of algorithms which is known as centric based clustering. The K-mean clustering takes K as an input parameter and divides the set of objects into n number of clusters such that the result should be high in intracluster similarity and low in intercluster similarity. In this, the examples are partitioned into various clusters in such a way that these clusters are optimal by following some criteria. The name has been derived from the various factors that will form the k clusters. In this cluster the center part is the arithmetic mean of all items encloses that cluster.

Advantages of K-mean Clustering

1. If we keep k smalls and variables are large, then many times K-mean is faster than the hierarchical clustering.

2. If the clusters are globalised in nature, K-Means produce tighter clusters than hierarchical clustering.

3. 3. It will best results when the data set are distinct.

4. It gives good results than classification.

5. It is fast and easy to understand.

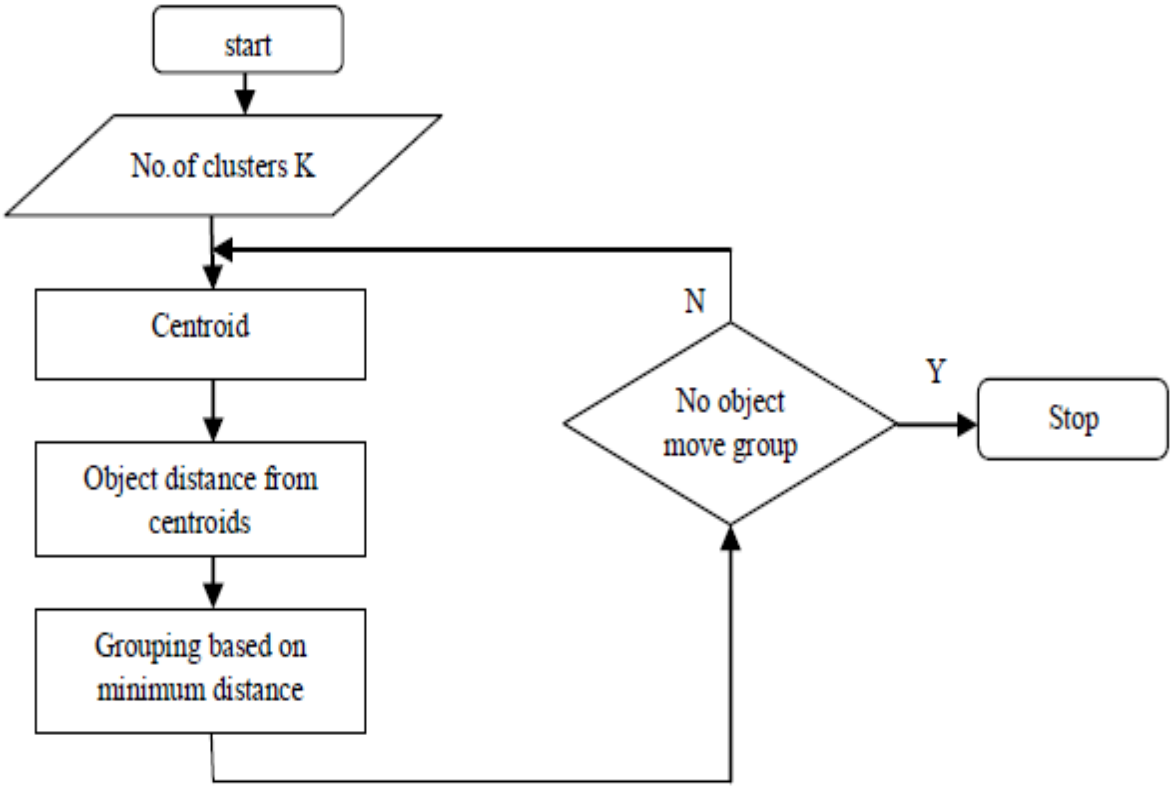


Figure 3.3 .K-means Flow Chart

C. Decision Tree

Decision tree is a influential and popular technique used in classification and prediction. There are a variety of algorithms used for making decision trees. Using any of the traditional algorithms such as C4.5, ID3 etc. construct a decision tree T from a set of training dataset. Decision tree is one of the mainly used data mining techniques. It is very model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. It is like hierarchal technique.

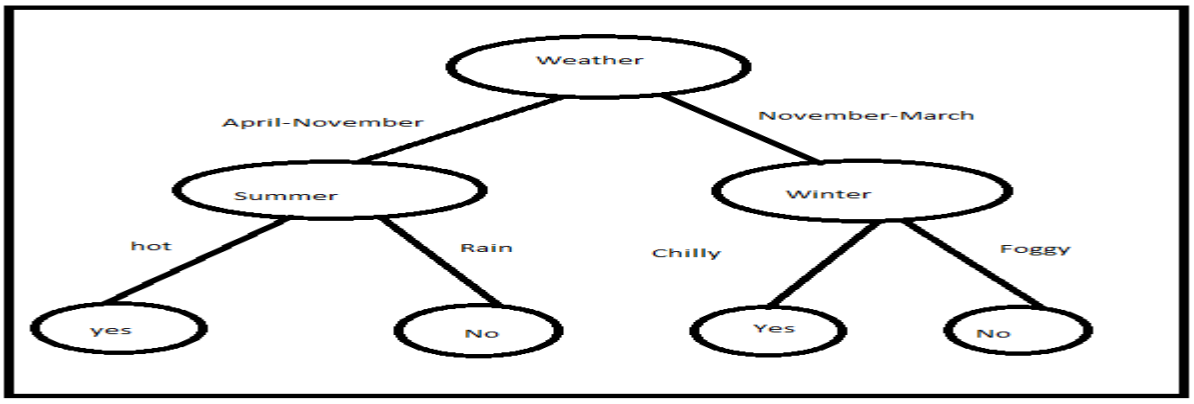


Figure 3.4 Decision Tree

3.5 PROPOSED SYSTEM ARCHITECTURE

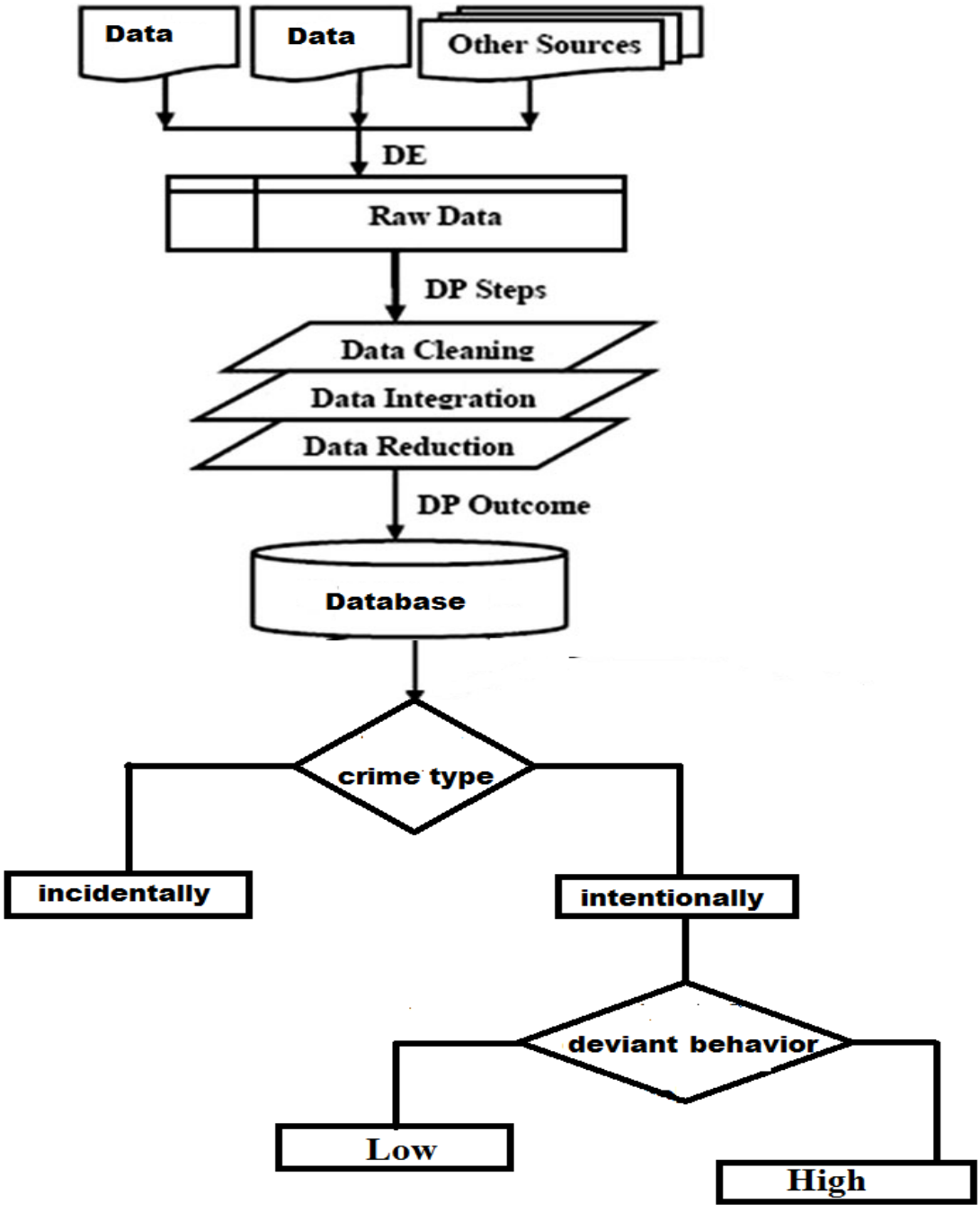


figure 3.5. The Research Methodology Work Flow.

This research mainly focused on an approach for clustering different type of criminals and classification of deviant behavior of juveniles of Indian correction centers using data mining techniques.

Our approach is divided into given modules, namely—data extraction (DE), data preprocessing (DP), clustering, and classification. First module, DE extracts the unstructured crime dataset from various crime Web sources.

Second module, DP cleans, integrates and reduces the extracted crime data into structured crime instances. The data can be collected from different sources and that data is noisy and incomplete data. Data extraction is used to make data into raw data.

Data pre-processing is used on raw data that is data cleaning, integration and reduction to make it pure data. To represent these instances using predefined crime attributes. Some modules are useful for identify crime type and deviant behavior prediction, respectively.

Criminal detection is analyzed using clustering K-means algorithm, which iteratively generates two crime clusters that are based on similar crime attributes. Cluster 0 means intentionally and cluster 1 means incidentally.

The deviant behavior of the criminal like violence level of the criminal can be predicted by using classification algorithm. The NEA algorithm is used as a classification algorithm. It is the enhanced algorithm. The valves used in analysis and prediction of deviant behavior of juvenile in correction centers. After that the classification the result is violence level is high and low.

From result, if the violence level is high than extra precautions is needed and if low than less precautions is needed.

The algorithm for the proposed system is as follows:

START

1. Collection of the real data.

2. Preprocess the collected data and make it in pure data.

3. Apply the k means clustering algorithm and the most wanted clusters are formed.

4. Now apply the attributes as input for classification.

5. The classification technique like decision tree calculates the values.

6. If the value is greater than threshold, go to 6 step else go to 7 step.

7. The criminal violence level is high/low.

8. If the value is greater than threshold then go to step 9,else to step 10.

9. High means criminal is harmful.

10. Generate the Alarm.

END

3.6 TOOLS OF DATA ANALYSIS

3.6.1 WEKA (Waikato Environment for Knowledge Analysis)

Weka (Waikato Environment for Knowledge Analysis) is a popular tool of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software accessible under the GNU General Public License. The algorithms can either be apply directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-matched for developing new machine learning schemes.

Originally this tool was designed for analyzing data from agriculture field but now a days it is used for research and education purpose.

Advantages of Weka include:

- This tools is free available under the GNU General Public License[12]
- portability- basically this tool is implemented in Java language but run on any current platform.
- a wide-range of data is preprocessing and modeling techniques are applied.
- Easy to use due to its graphical user interface.

The Explorer [13] interface features several techniques used for providing access to the main components of the workbench:

- The Preprocess has feature to import data from system or database in the CSV file. After importing data is preprocessed by using the filter algorithm and filtering can be supervised or unsupervised. The filtering technique is used to transform data from one form to another like nominal to numeric.
- The Classify provides facility to user to use classification and regression techniques. It is used to predict the accuracy, result prediction and also the visualizing of prediction results, visualization are like ROC, curves, cost estimation and trees formation.

- The Associate [14] is used for identifying the important and invisible interrelationship between the attributes and can be access by association rule learners.

- The Cluster provides the access to the clustering techniques like EM, the simple k-means algorithm. It is used for implementation of a mixture of normal distribution. It is also an expectation algorithm.

- The Select attributes is used for finding the most predictive attributes in a dataset.

- The Visualize feature is used to show cost estimation, curves in graph format in x and y axis. It shows plot matrix and scatter plot can compress, select and enlarge by using various selection features.

3.6.2 NetBeans

NetBeans is a platform used for developing software which is written in Java. The NetBeans provide a Platform which allowed user to develop an applications from a set of modular software components called *modules*. [6] The NetBeans IDE is mainly used in Java, but it also supports other different languages, like PHP, C/C++and HTML5.

The NetBeans Team dynamically supports the product and also looks for suggestions regarding feature from the whole society. Every release is for the society testing and feedback.

The two basic products, the NetBeans IDE and NetBeans Platform, are free for commercial and non-commercial use.[7] The source code of both product is available to anyone, within the terms of use. The legal section contains information regarding licensing, copyright issues, privacy policy and terms of use.

CHAPTER 4

RESULTS & DISCUSSIONS

4.1 IMPLEMENTATION DETAILS

The proposed research is implemented for detection of crime reason and the violence level and this research has three phases, first preprocessing of the data second making clusters by using trained dataset through clustering algorithm and third classification and result analysis. Experiments are conducted on real world data. The collected data is distributed into two categories, first is intentionally that cluster 0 and second is incidentally cluster 1 are used for training and Cluster 0 Dataset is further used for classification. The various implementation steps are given below.

Step1: Collection of data

The data is collected from various 11 police stations (divisions) of Jalandhar. Not whole but they provide essential information about the juvenile criminals as per rules and permission. The information is shown in the table.

Table 2: Data used for training

	A	B	C	D	E	F	G	H
1	**Crime_Type**	**Month**	**Year**	**Crime_Cc**	**Age**	**Divisions**	**District**	**Location**
2	Murder	1	2012	1	15	D1	Jalandhar	Home
3	Attempt to commit murder	2	2012	1	18	D2	Jalandhar	office
4	Theft	5	2012	2	12	D3	Jalandhar	School
5	Attempt to commit murder	9	2012	1	15	D4	Jalandhar	School
6	ROBBERY	10	2012	1	12	D5	Jalandhar	Road
7	BURGLARY	12	2011	1	14	D6	Jalandhar	Shop
8	Theft	12	2011	1	16	D7	Jalandhar	Shop
9	RIOTS	9	2012	1	15	D8	Jalandhar	Road
10	ARSON	2	2012	1	13	D9	Jalandhar	Bus stand
11	Rape	5	2012	1	18	D5	Jalandhar	Shop
12	ROBBERY	5	2012	1	18	D9	Jalandhar	Station
13	Rape	9	2012	1	15	D11	Jalandhar	Home
14	Murder	10	2012	1	12	D10	Jalandhar	office
15	Attempt to commit murder	12	2011	1	14	D6	Jalandhar	School
16	RAPE	12	2011	1	16	D6	Jalandhar	School
17	KIDNAPPING & ABDUCTION	9	2012	1	15	D7	Jalandhar	Road
18	KIDNAPPING & ABDUCTION	2	2012	1	13	D8	Jalandhar	Shop
19	SEXUAL HARASSMENT	12	2011	1	14	D8	Jalandhar	Shop
20	theft	12	2011	1	16	D2	Jalandhar	Road
21	KIDNAPPING & ABDUCTION	9	2012	1	15	D2	Jalandhar	Bus stand
22	ARSON	2	2012	1	13	D2	Jalandhar	Shop

Step2: Data Transformation

The next step is the data transformation .In this, the data is transformed according to the desires and policy to obtain the hidden patterns and to find out the unknown relationship among the data. In this research to obtain the unknown pattern from the data , WEKA is used as a tool. In Weka (Waikato Environment for Knowledge Analysis) the algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Step3: Data File uploaded

The first step after the data transformation is cluster the data into two clusters using K-means clustering algorithm. After executing K-means algorithm on the dataset transactions in Table, the clusters were formed cluster 0 and cluster 1.

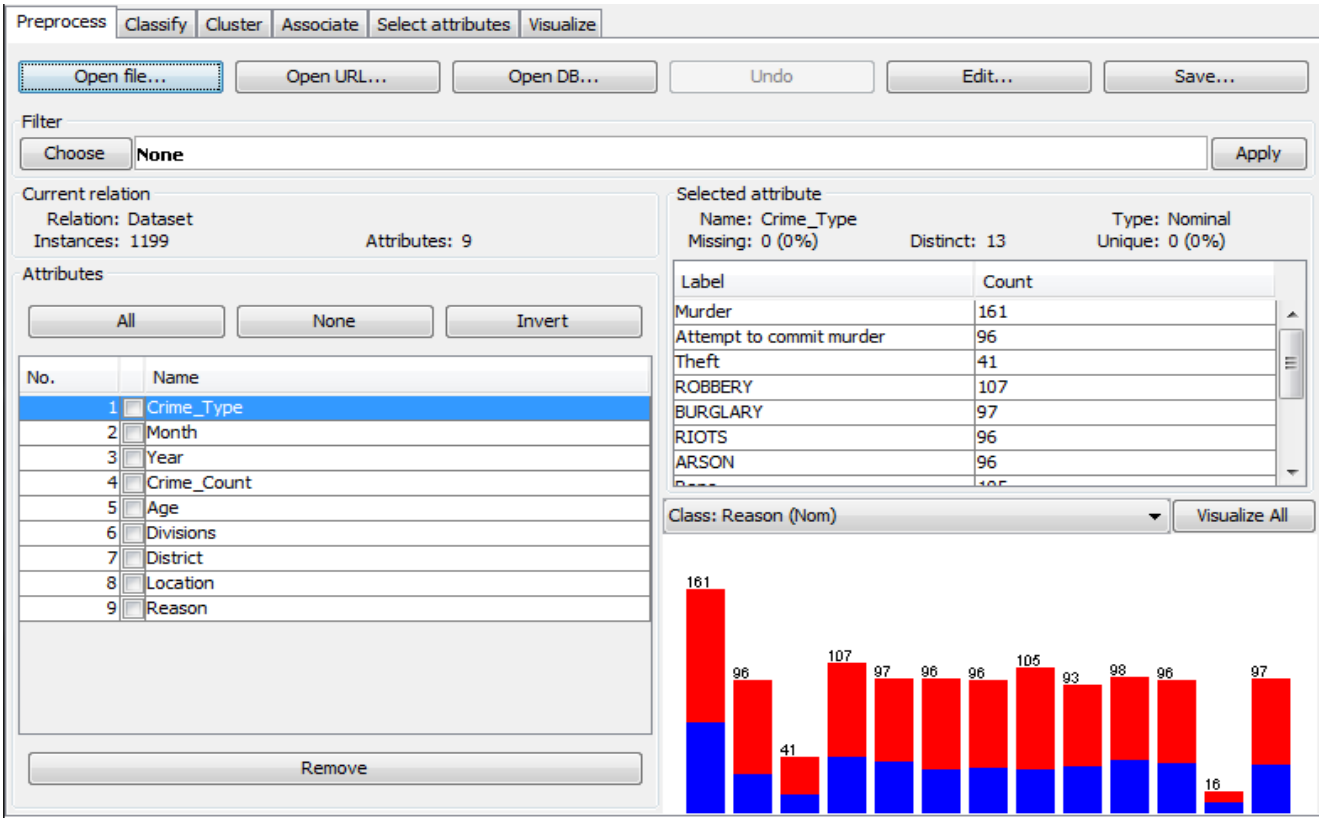


Figure 4.1: Data File Uploaded

Step4: Use of clustering algorithm

When the file is data file is uploded then the filter is choosen from the drop down menu. the clustering is performed. In this the K-means clustering is used. The figure shows the selection of the clustering algorithm. Now on click of the start button, the clustering is being performed and we get the results.The figure shown below display the results of clustering.It included the eight attributes and two clusters cluster 0 and cluster 1 is formed.

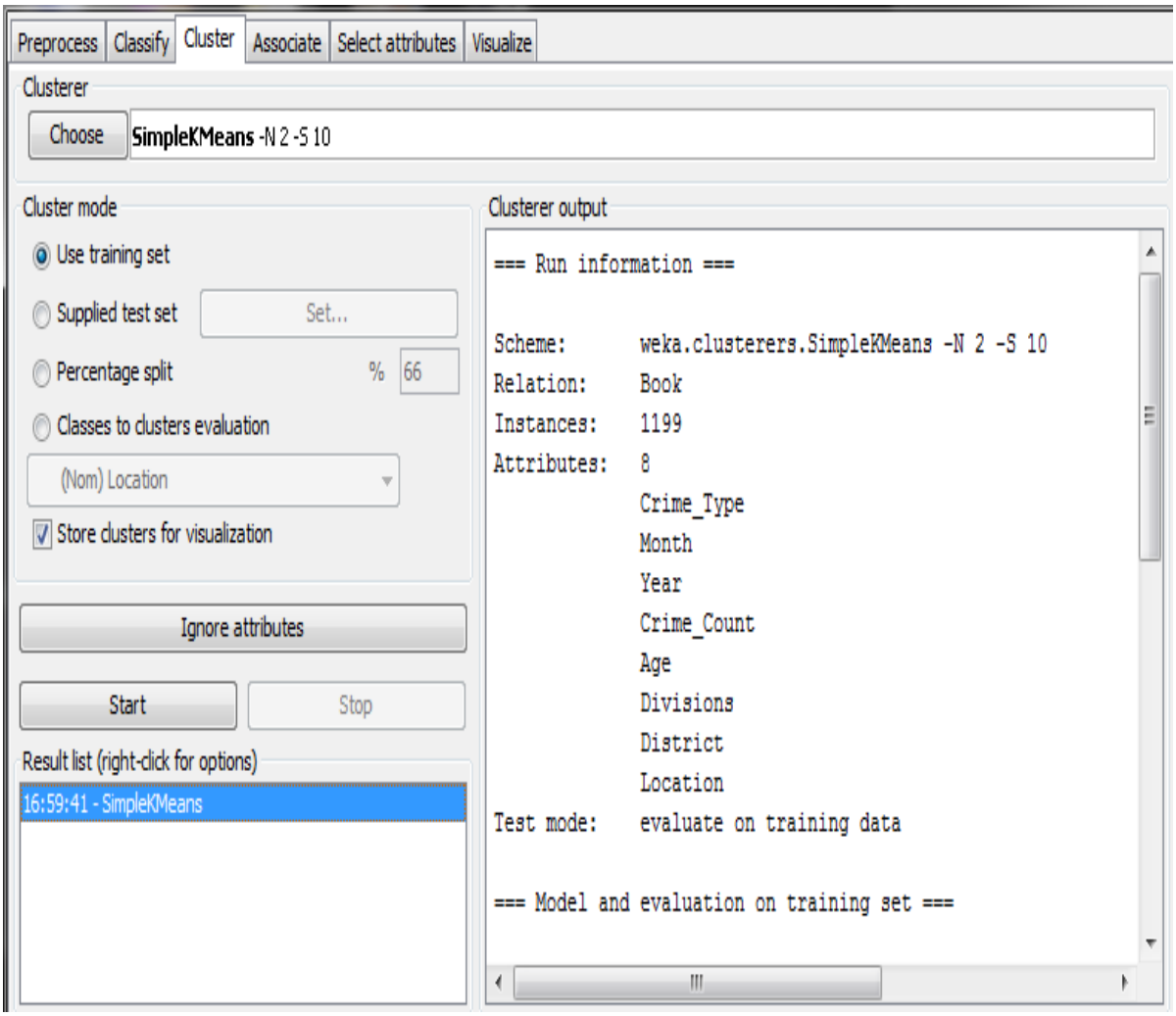


Figure 4.2: Cluster Performance.

Four iteration were used in formulation of the two clusters and within the clusters the sum of squared error is 3549.8.The missing values in the data set is filled by the mean/mode between the data values the exact formation of the clusters and the centorid of the two clusters. Cluster 0 is the intentionally cluster which contain the criminal data and it contain 64%of the total data set and cluster 1 is the incidentally cluster which contains the innocent criminal data set .

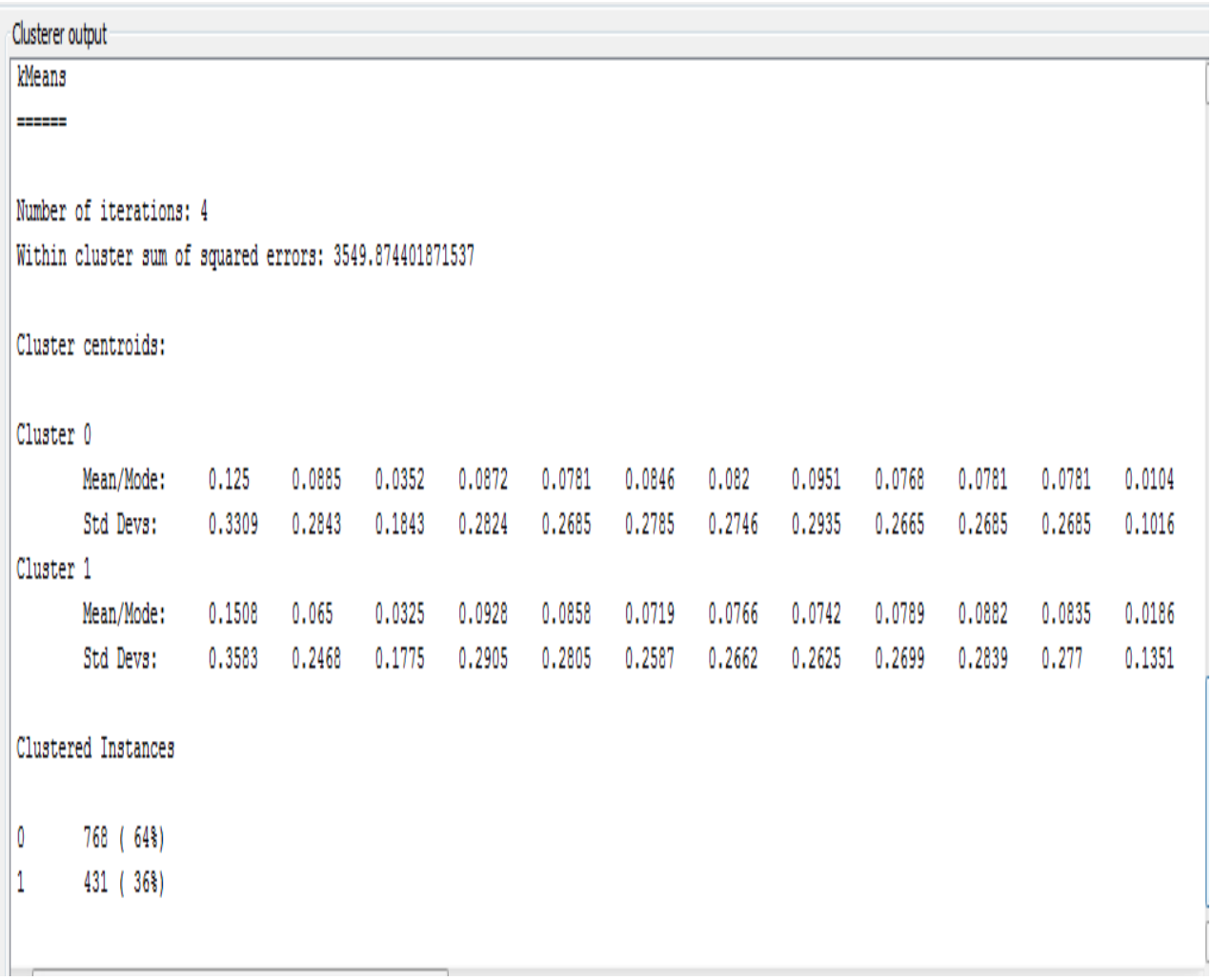


Figure 4.3: Cluster formulation

The pictorial representation of the cluster is shown in figure 4.4 .The blue color represent the data that fit in to cluter 0 and the other colour represent the data fit into cluster 1.The same can be represented in the form of plot graph as shown in figure 4.5..

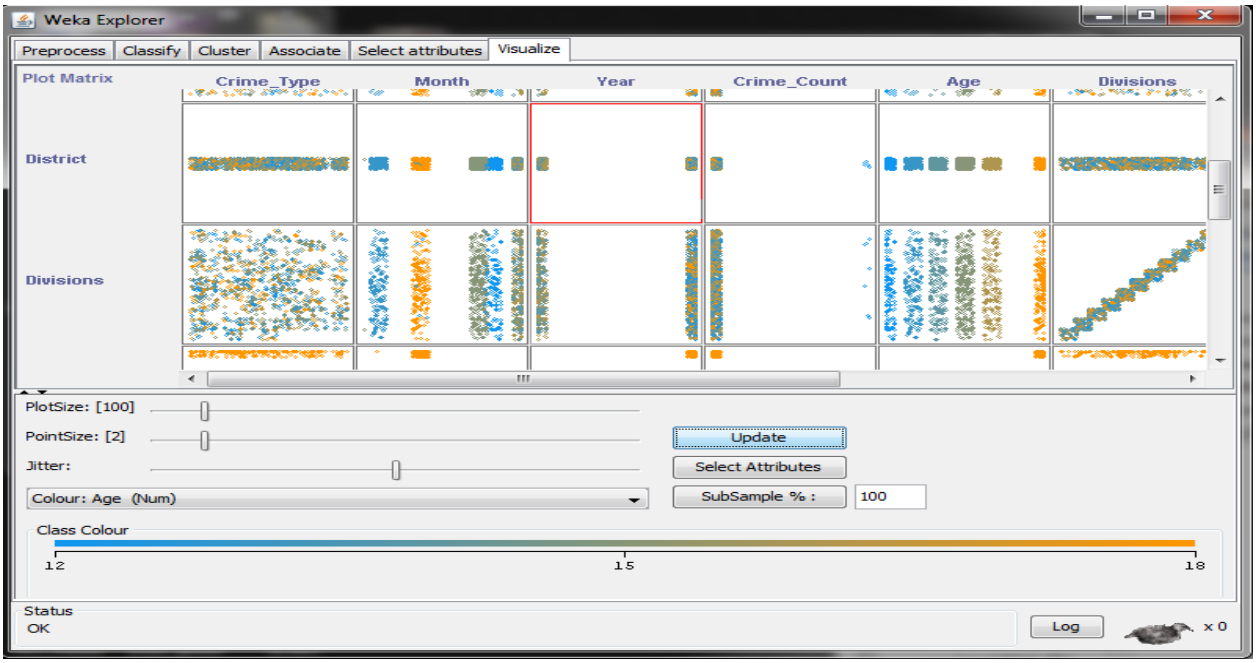


Figure4.4: Visualization of the Cluster

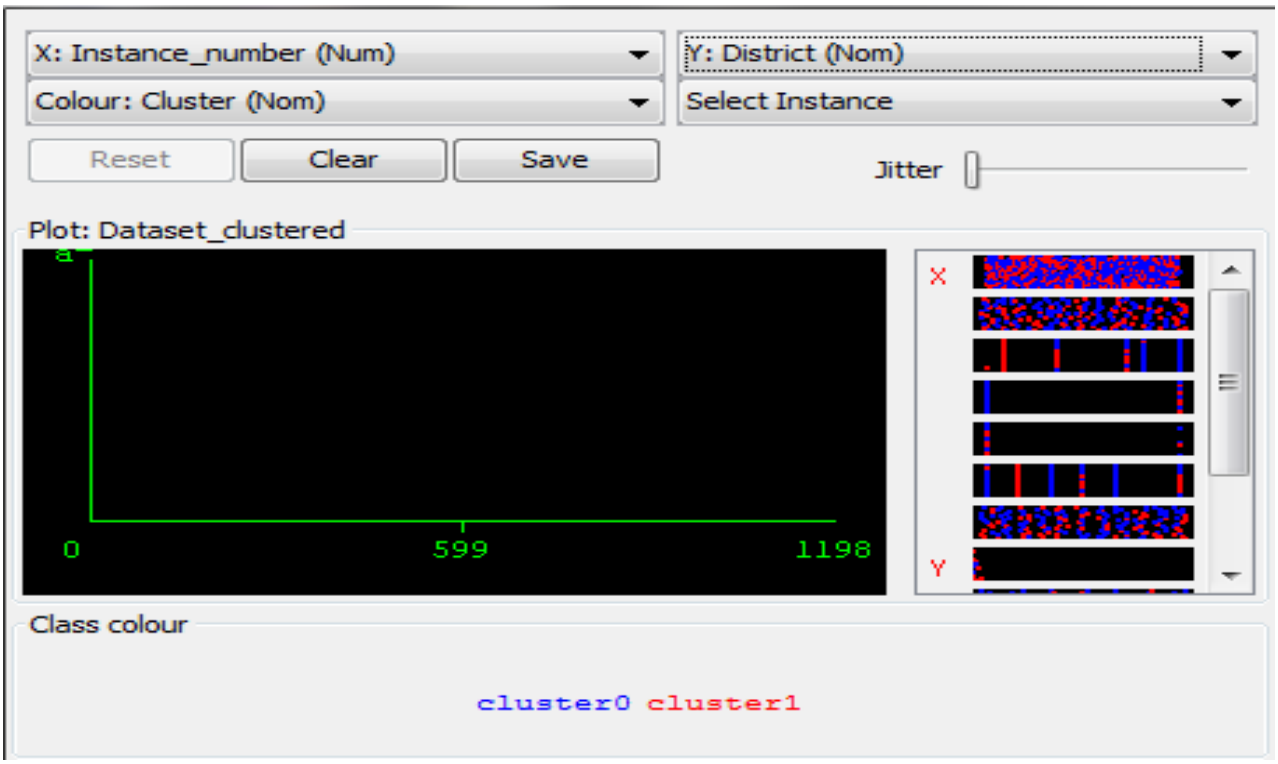


Figure 4.5: Plot graph for the Cluster Assignment

The first step towards performing clustering is that the arff file is uploaded in the weka tool and the attributes are selected which are required other attributes are ignored.

```
0,Murder,1,2012,1,15,D1,Jalandhar,Home,cluster0
1,'Attempt to commit murder',2,2012,1,18,D2,Jalandhar,office,cluster0
2,Theft,5,2012,2,12,D3,Jalandhar,School,cluster0
3,'Attempt to commit murder',9,2012,1,15,D4,Jalandhar,School,cluster0
4,ROBBERY,10,2012,1,12,D5,Jalandhar,Road,cluster1
5,BURGLARY,12,2011,1,14,D6,Jalandhar,Shop,cluster0
6,Theft,12,2011,1,16,D7,Jalandhar,Shop,cluster0
7,RIOTS,9,2012,1,15,D8,Jalandhar,Road,cluster1
8,ARSON,2,2012,1,13,D9,Jalandhar,'Bus stand',cluster0
9,Rape,5,2012,1,18,D5,Jalandhar,Shop,cluster0
10,ROBBERY,5,2012,1,18,D9,Jalandhar,Station,cluster0
11,Rape,9,2012,1,15,D11,Jalandhar,Home,cluster0
12,Murder,10,2012,1,12,D10,Jalandhar,office,cluster0
13,'Attempt to commit murder',12,2011,1,14,D6,Jalandhar,School,cluster1
14,RAPE,12,2011,1,16,D6,Jalandhar,School,cluster1
15,'KIDNAPPING & ABDUCTION',9,2012,1,15,D7,Jalandhar,Road,cluster1
16,'KIDNAPPING & ABDUCTION',2,2012,1,13,D8,Jalandhar,Shop,cluster0
17,'SEXUAL HARASSMENT',12,2011,1,14,D8,Jalandhar,Shop,cluster0
18,theft,12,2011,1,16,D2,Jalandhar,Road,cluster1
19,'KIDNAPPING & ABDUCTION',9,2012,1,15,D2,Jalandhar,'Bus stand',cluste
20,ARSON,2,2012,1,13,D2,Jalandhar,Shop,cluster0
21,Rape,5,2012,1,18,D4,Jalandhar,Station,cluster0
22,MOLESTATION,5,2012,1,18,D4,Jalandhar,Home,cluster0
23,MOLESTATION,9,2012,1,15,D5,Jalandhar,office,cluster1
24,Murder,10,2012,1,12,D6,Jalandhar,School,cluster0
25,'Attempt to commit murder',12,2011,1,14,D7,Jalandhar,School,cluster1
26,RAPE,9,2012,1,15,D8,Jalandhar,Road,cluster1
27,'KIDNAPPING & ABDUCTION',10,2012,1,12,D7,Jalandhar,Shop,cluster0
28,ROBBERY,12,2011,1,14,D5,Jalandhar,Shop,cluster1
```

Figure 4.6: Clustering Of Each dataset.

The table 3 shows the sample data grouping of the table 2. From the sample data as shown in table 3, 763 data belongs to cluster0 and 431 belongs to cluster1.

Table 3: Result of Clustering

	A	B	C	D	E	F	G	H	I	J
1	Crime_Ty	Month	Year	Crime_Co	Age	Divisions	District	Location	Reason	Cluster
2	Murder	1	2012	1	15	D1	Jalandhar	Home	incidentally	Cluster 1
3	Attempt t	2	2012	1	18	D2	Jalandhar	office	intentionally	Cluster 0
4	Theft	5	2012	2	12	D3	Jalandhar	School	intentionally	Cluster 0
5	Attempt t	9	2012	1	15	D4	Jalandhar	School	intentionally	Cluster 0
6	ROBBERY	10	2012	1	12	D5	Jalandhar	Road	intentionally	Cluster 0
7	BURGLARY	12	2011	1	14	D6	Jalandhar	Shop	intentionally	Cluster 0
8	Theft	12	2011	1	16	D7	Jalandhar	Shop	intentionally	Cluster 0
9	RIOTS	9	2012	1	15	D8	Jalandhar	Road	intentionally	Cluster 0
10	ARSON	2	2012	1	13	D9	Jalandhar	Bus stand	incidentally	Cluster 1
11	Rape	5	2012	1	18	D5	Jalandhar	Shop	intentionally	Cluster 0
12	ROBBERY	5	2012	1	18	D9	Jalandhar	Station	incidentally	Cluster 1
13	Rape	9	2012	1	15	D11	Jalandhar	Home	intentionally	Cluster 0
14	Murder	10	2012	1	12	D10	Jalandhar	office	incidentally	Cluster 1
15	Attempt t	12	2011	1	14	D6	Jalandhar	School	intentionally	Cluster 0
16	RAPE	12	2011	1	16	D6	Jalandhar	School	intentionally	Cluster 0

STEP 5: Training of model using Classification

Now, when the complete data is clustered into two groups and in order to reduce the dataset and for the detection of the intentional criminal we now pay attention on the datasets or users belonging to the cluster 0. Now, train the model using the classification C4.3 algorithm and NEA algorithm. The table 4 shows the input data provided to the classifier for the training purpose. It includes the criminal who belonged to cluster 0.

Table 4: Input Data for Classification

	A	B	C	D	E	F	G	H	I
1	Crime_Type	Month	Year	Crime_Count	Age	Divisions	District	Location	Cluster
2	Murder	1	2012	1	15	D1	Jalandhar	Home	Cluster 0
3	Attempt to commit murder	2	2012	1	18	D2	Jalandhar	office	Cluster 0
4	Theft	5	2012	2	12	D3	Jalandhar	School	Cluster 0
5	Attempt to commit murder	9	2012	1	15	D4	Jalandhar	School	Cluster 0
6	BURGLARY	12	2011	1	14	D6	Jalandhar	Shop	Cluster 0
7	Theft	12	2011	1	16	D7	Jalandhar	Shop	Cluster 0
8	ARSON	2	2012	1	13	D9	Jalandhar	Bus stand	Cluster 0
9	Rape	5	2012	1	18	D5	Jalandhar	Shop	Cluster 0
10	ROBBERY	5	2012	1	18	D9	Jalandhar	Station	Cluster 0
11	Rape	9	2012	1	15	D11	Jalandhar	Home	Cluster 0
12	Murder	10	2012	1	12	D10	Jalandhar	office	Cluster 0
13	KIDNAPPING & ABDUCTION	2	2012	1	13	D8	Jalandhar	Shop	Cluster 0
14	SEXUAL HARASSMENT	12	2011	1	14	D8	Jalandhar	Shop	Cluster 0
15	KIDNAPPING & ABDUCTION	9	2012	1	15	D2	Jalandhar	Bus stand	Cluster 0
16	ARSON	2	2012	1	13	D2	Jalandhar	Shop	Cluster 0
17	Rape	5	2012	1	18	D4	Jalandhar	Station	Cluster 0
18	MOLESTATION	5	2012	1	18	D4	Jalandhar	Home	Cluster 0
19	Murder	10	2012	1	12	D6	Jalandhar	School	Cluster 0
20	KIDNAPPING & ABDUCTION	10	2012	1	12	D7	Jalandhar	Shop	Cluster 0
21	Theft	2	2012	1	13	D3	Jalandhar	Shop	Cluster 0
22	Theft	5	2012	1	18	D9	Jalandhar	Station	Cluster 0

Now when the input is fed into the weka tool it includes the nine mentioned attributes and the clustered values obtained from the clustering. From the figure it is clear that the data set has no missing values and 13 distinct. All the 763 instances are unique and distinct. No value is repeated.

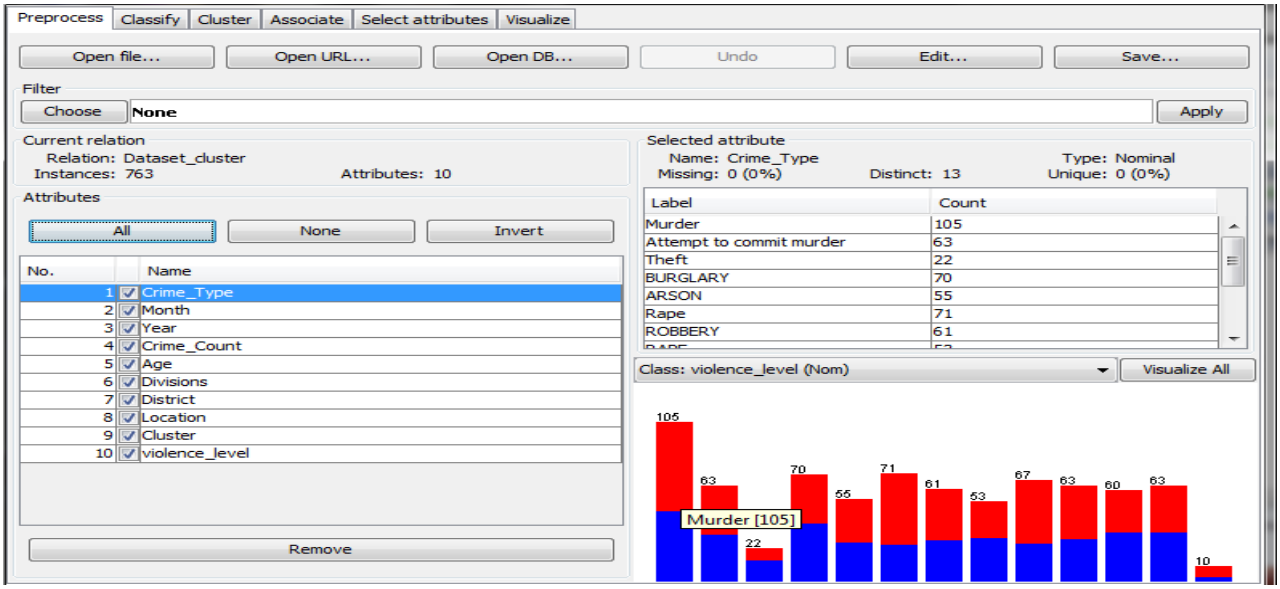


Figure 4.7: Clustered result as input to Classification

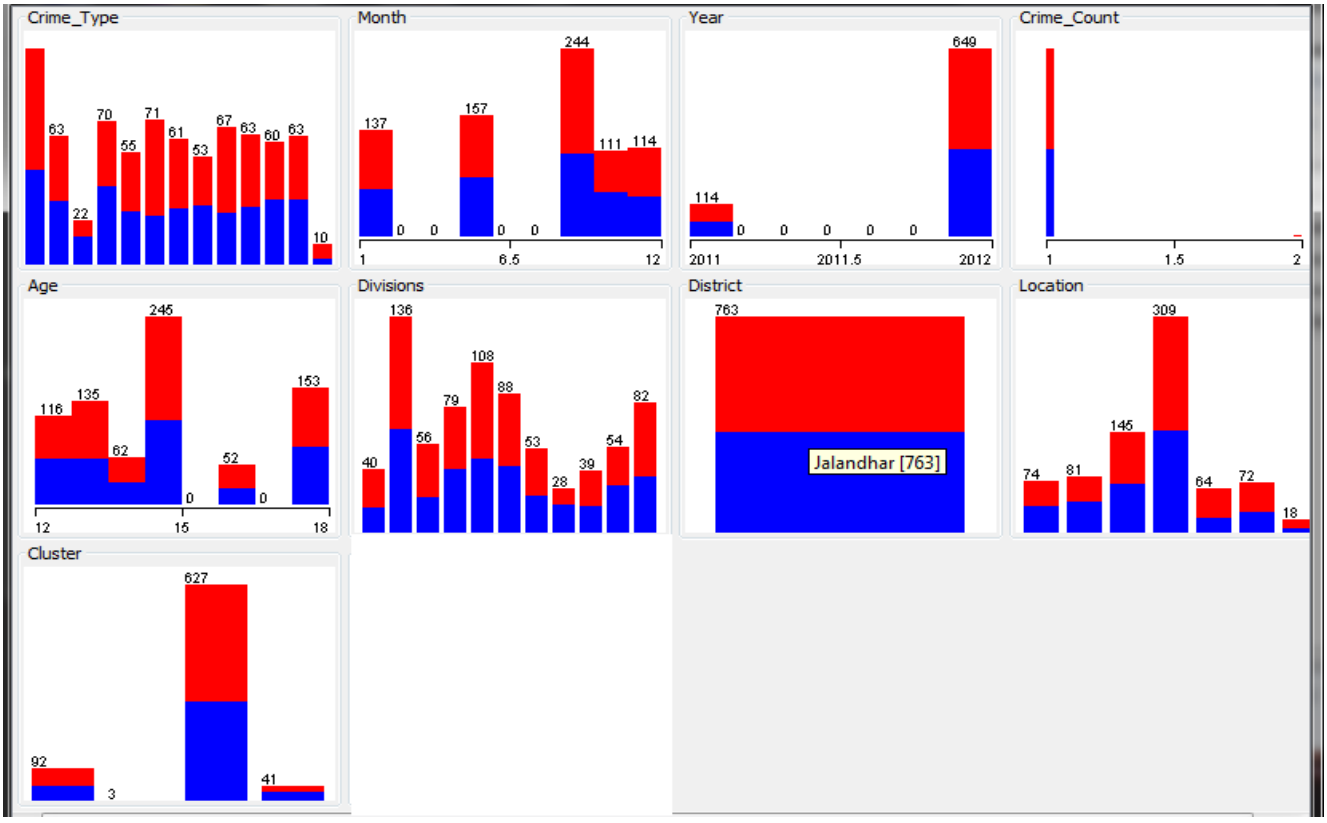


Figure 4.8: Visualization of Clustered Output

In the above figure4.8, the visualization of different attributes with values and graphical representation of Clustered output data.

In the below figures 4.9, 4.10, the classifier output there is 763 instances and 9 attributes. The time taken to build a model is 0.02 seconds. The correctly classified instances 408 (53.4%). The confusion matrix classified as low and high.

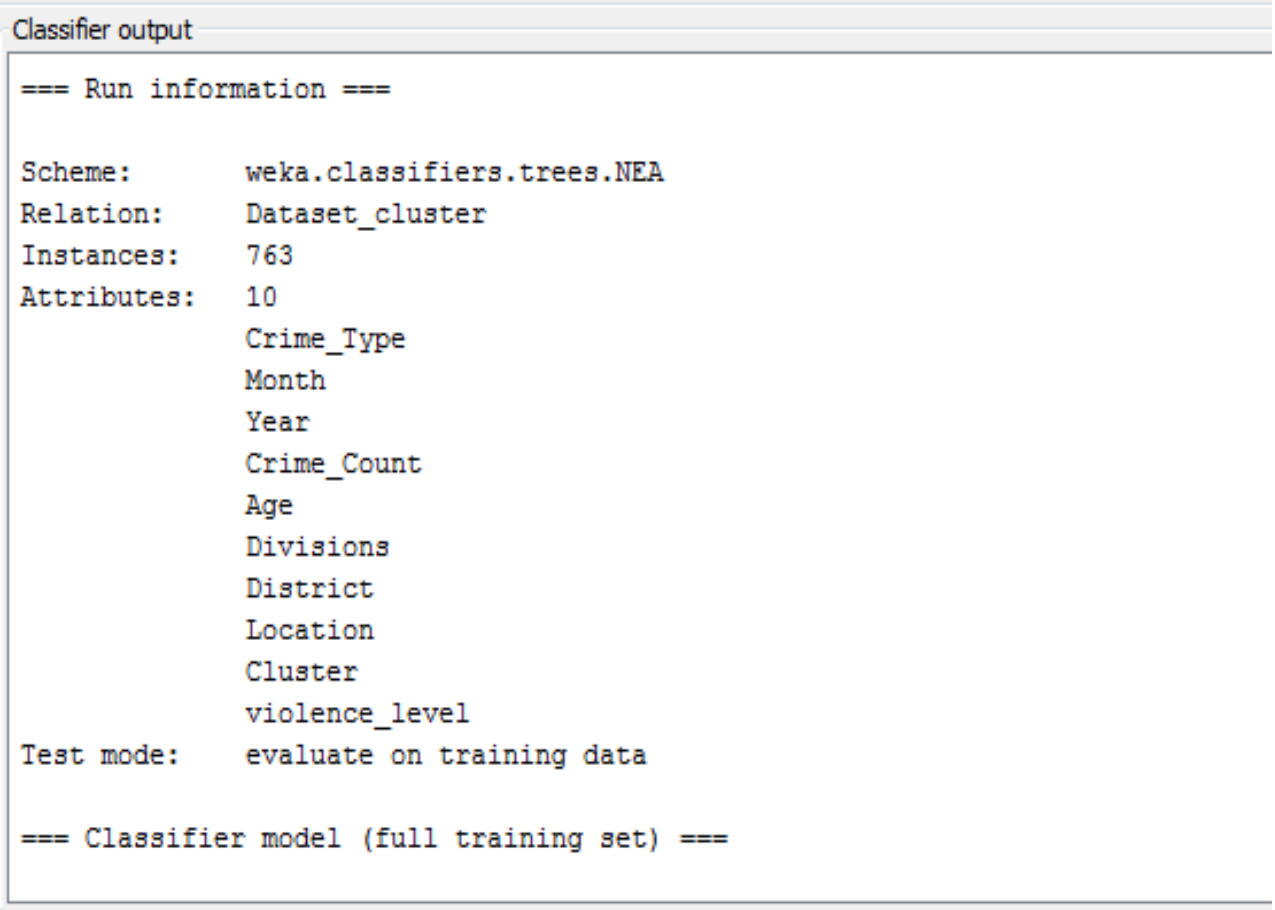


Figure 4.9: The Classifier Output

```
Classifier output
Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         408               53.4731 %
Incorrectly Classified Instances       355               46.5269 %
Kappa statistic                          0
Mean absolute error                      0.4943
Root mean squared error                  0.4971
Relative absolute error                 99.3311 %
Root relative squared error             99.6656 %
Total Number of Instances              763

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0          0          0          0          0        low
  1          1          0.535      1          0.697    high

=== Confusion Matrix ===

   a   b   <-- classified as
   0 355 |   a = low
   0 408 |   b = high
```

Figure 4.10: The Classifier Output

4.2 RESULTS

In every model, the accuracy and the time plays a important role in the acceptance of that model for the application. It's applicable for the given proposed system as well. Table 5 shows the main result of the implementation

Table 5: Result.

	A	B	C	D	E	F	G	H	I	J
1	Crime_Type	Month	Year	Crime_Count	Age	Divisions	District	Location	Cluster	violence_level
2	Murder	1	2012	1	15	D1	Jalandhar	Home	Cluster 0	low
3	Attempt to commit murder	2	2012	1	18	D2	Jalandhar	office	Cluster 0	low
4	Theft	5	2012	2	12	D3	Jalandhar	School	Cluster 0	low
5	Attempt to commit murder	9	2012	1	15	D4	Jalandhar	School	Cluster 0	high
6	BURGLARY	12	2011	1	14	D6	Jalandhar	Shop	Cluster 0	high
7	Theft	12	2011	1	16	D7	Jalandhar	Shop	Cluster 0	low
8	ARSON	2	2012	1	13	D9	Jalandhar	Bus stand	Cluster 0	high
9	Rape	5	2012	1	18	D5	Jalandhar	Shop	Cluster 0	high
10	ROBBERY	5	2012	1	18	D9	Jalandhar	Station	Cluster 0	low
11	Rape	9	2012	1	15	D11	Jalandhar	Home	Cluster 0	high
12	Murder	10	2012	1	12	D10	Jalandhar	office	Cluster 0	low
13	Attempt to commit murder	12	2011	1	14	D6	Jalandhar	School	Cluster 0	high
14	RAPE	12	2011	1	16	D6	Jalandhar	School	Cluster 0	high
15	KIDNAPPING & ABDUCTION	2	2012	1	13	D8	Jalandhar	Shop	Cluster 0	high
16	SEXUAL HARASSMENT	12	2011	1	14	D8	Jalandhar	Shop	Cluster 0	low
17	KIDNAPPING & ABDUCTION	9	2012	1	15	D2	Jalandhar	Bus stand	Cluster 0	low
18	ARSON	2	2012	1	13	D2	Jalandhar	Shop	Cluster 0	high
19	Rape	5	2012	1	18	D4	Jalandhar	Station	Cluster 0	low
20	MOLESTATION	5	2012	1	18	D4	Jalandhar	Home	Cluster 0	high
21	Murder	10	2012	1	12	D6	Jalandhar	School	Cluster 0	low
22	KIDNAPPING & ABDUCTION	10	2012	1	12	D7	Jalandhar	Shop	Cluster 0	high
23	Theft	2	2012	1	13	D3	Jalandhar	Shop	Cluster 0	low
24	Theft	5	2012	1	18	D9	Jalandhar	Station	Cluster 0	high
25	Theft	5	2012	1	18	D10	Jalandhar	Home	Cluster 0	high

Thus it can be easily concluded from the below figure that the proposed model has a less error rate and need less time for model. The error rate of NEA is 99.4 and time taken 0.02.

Figure 4.11: Result comparison of C4.5 and NEA.

Chapter 5
CONCLUSION AND FUTURE SCOPE

The research focuses on an analysis the type of crime and prediction of deviant behaviour of the juvenile in the correction centre. The data can be collected from different sources and that data is noisy and incomplete data. Data extraction is used to make data into raw data. Data pre-processing is used on raw data that is data cleaning, integration and reduction to make it pure data. The pure database gives more accuracy and efficiency. The objectives of the research will be achieved like data extraction, data preprocessing of incomplete and inconsistent data, accurately and efficiently analyze the growing volume of data, explore and enhance clustering and classification algorithm to identify type of crime, and construct an efficient framework for predicting deviant behavior of juveniles. The time consumption and storage space can be decreases. The categorization of crime type can be done by using algorithm i.e. crime is incidentally and intentionally. The categorization of deviant behaviour can be done by using classification algorithm i.e. the juvenile is harmful or not means violence level is low or high. The taken to build a model is 0.02 and the error rate is 99.4%.

.

5.2 FUTURE SCOPE

This research focused on reason that the crime is done intentionally and incidentally and the deviant behaviour means the violence level is low or high. In future the accuracy rate can be improved.

References

Reference from a book:

[1]. Han J and Kamber M, Data Mining: Concepts and Techniques (3rd ed.). *Morgan Kaufmann, San Francisco*, CA, 2012

[2]. Ben Krose,Patrick van der smagt, "An introduction to Neural Networks" ,1996.

[3]. David Kriesel, " A Brief Introduction to Neural Networks",2011

Reference from an article:

[4]. Kaur Sukhdeep ,(2013) "Punjab's case on juvenile crime: no age of innocence", Hindustan Times, Chandigarh Updated: Sep 25, 17:32 IST

Reference from web page:

[5]. https://www.ncjrs.gov/html/ojjdp/218587/images/p10-1.gif

[6]. http://en.wikipedia.org/wiki/NetBeans

[7]. https://netbeans.org/about/

[8]. http://journalistsresource.org/wp-content/uploads/2014/10/Trends-in-juvenile-arrests-202-2011-DOJ.jpgtable

[9]. http:// ncrb.nic.co.in/

[10]. http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#The%20Learning%20Process

[11]. http://www.dataminingarticles.com/data-mining-introduction.html

[12]. http://lifehacker.com/5237503/five-best-free-data-recovery-tools

[13].]http://sourceforge.net/projects/weka/

[14]. http://wikipedia.org/weka.php

[15]. Shehata Shady, Karray Fakhri, and Kamel S Mohamed, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, october 2010.

[16]. Leung Sak Kwong, Lee Hong Kin, Wang Feng Jin, "Data Mining on DNA Sequences of Hepatitis B Virus", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 2, march/april 2011.

[17]. Mansoori G.Eghbal, "FRBC: A Fuzzy Rule-Based Clustering Algorithm", IEEE Transactions on Fuzzy Systems, vol. 19, no. 5, october 2011.

[18]. Malathi. A, Dr.Baboo Santhosh S, " An Enhanced Algorithm to Predict a Future Crime using Data Mining", *International Journal of Computer Applications (0975 – 8887)* Volume 21– No.1, May 2011.

[19]. Wang Lijun, Rege Manjeet, Dong Ming, Member, IEEE, and Yongsheng Ding, Senior Member, IEEE, "Low-Rank Kernel Matrix Factorization for Large-Scale Evolutionary Clustering" , *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, june 2012.

[20]. Vongsingthong Suwimon, Wisitpongphan Nawaporn, "classification of university students behavior in sharing information on facebook*", IEEE/International Joint Conference on Computer Science and Software Engineering*, 2014.

[21]. Tayal Kumar Devendra , Jain Arti , Arora Surbhi ,AgarwalSurbhi, Gupta Tushar, Tyagi Nikhil, "Crime detection and criminal identification in India using data mining Techniques*", Springer*, march/april 2014.

[22]. Wei Wei, Jinjiu Li, Longbing Cao, Yuming Ou, Jiahang Chen , "Effective detection of sophisticated online banking fraud on extremely imbalanced data", *Springer Science Bussiness Media, LLC 2012*.

[23]. Mansingh Gunjan, Rao Lila, Osei-Bryson Kweku-Muata,"Profling internet banking users: A Knowledge discovery in data mining process model process model based approach", *Springer Science Business Media New York 2013*.

Appendix

Glossary of Terms

analytical model	An organization, development and analyzing a dataset. For example, a decision tree is a model for the classification of a dataset.
Artificial neural networks	Non-linear predictive models that learn through instruction and be similar to biological neural networks in structure.
CART	A decision tree technique used for classification of a dataset. Provides a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome.
CHAID	Chi Square Automatic Interaction Detection. A decision tree technique used for classification of a dataset. Provides a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome.
Classification	Classification is the most commonly used data mining technique, where a model is constructing to predict class. It is used to classify each item in a set of data into one of predefined set of classes or groups.
clustering	Clustering is a data mining technique that makes useful cluster of objects that have similar characteristic.
Crime type	Different type of crimes present in the constitution of law like murder, molestation, theft etc.

data cleansing	The process of cleaning that all values in a dataset are consistent and correctly recorded.
data mining	The extraction of hidden information from large amount of databases.
data navigation	The process of visualization of different dimensions, and levels of detail of a multidimensional database. See OLAP.
data visualization	The visual understanding of complex relationships in multidimensional data.
decision tree	A tree-shaped structure that represents a set of decisions. These decisions generate rules for the classification of a dataset. See CART and CHAID.
Deviant Behavior	The behavior which is not accepted in the society, which is against the law.
Exploratory data analysis	The use of graphic statistical techniques to learn about the structure of a dataset.
logistic regression	A linear regression that predicts the proportions of a categorical target variable, such as type of customer, in a population.
multidimensional database	A database designed for on-line analytical processing. Structured as a multidimensional hypercube with one axis per dimension.
multiprocessor computer	A computer that includes multiple processors connected by a network like parallel processing.

nearest neighbor	A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called a k-nearest neighbor technique.
non-linear model	An analytical model that does not assume linear relationships in the coefficients of the variables being studied.
OLAP	On-line analytical processing. Refers to array-oriented database applications that allow users to view, navigate through, manipulate, and analyze multidimensional databases.
parallel processing	The coordinated use of multiple processors to perform computational tasks. Parallel processing can occur on a multiprocessor computer or on a network of workstations or PCs.
predictive model	A structure and process for predicting the values of specified variables in a dataset.
Prospective data analysis	Data analysis that predicts future trends, behaviors, or events based on historical data.
RAID	Redundant Array of Inexpensive Disks. A technology for the efficient parallel storage of data for high-performance computer systems.
Retrospective data analysis	Data analysis that provides insights into trends, behaviors, or events that have already occurred.

rule induction	The extraction of useful if-then rules from data based on statistical significance.
SMP	Symmetric multiprocessor. A type of multiprocessor computer in which memory is shared among the processors.
Time series analysis	The analysis of a sequence of measurements made at specified time intervals. Time is usually the dominating dimension of the data.