

**DSpace Institution**

**DSpace Repository**

<http://dspace.org>

---

Information Technology

thesis

---

2023-11

# DEVELOPING A PREDICTIVE MODEL FOR RAPE CRIME PREVENTION USING DATA MINING TECHNIQUES

GEDAM, ESUBALEW MITIKU

---

<http://ir.bdu.edu.et/handle/123456789/15689>

*Downloaded from DSpace Repository, DSpace Institution's institutional repository*



**BAHIR DAR UNIVERSITY**

**BAHIR DAR INSTITUTE OF TECHNOLOGY**

**SCHOOL OF GRADUATE STUDIES**

**FACULTY OF COMPUTING**

**DEPARTMENT OF INFORMATION TECNOLOGY**

**MSc THESIS ON:-**

**DEVELOPING A PREDICTIVE MODEL FOR RAPE CRIME  
PREVENTION USING DATA MINING TECHNIQUES**

**BY:**

**GEDAM ESUBALEWMITIKU**

**NOVEMBER, 2023**

**BAHIR DAR, ETHIOPIA**



**BAHIR DAR UNIVERSITY**

**BAHIR DAR INSTITUTE OF TECHNOLOGY**

**FACULTY OF COMPUTING**

**DEVELOPING A PREDICTIVE MODEL FOR RAPE CRIME PREVENTION USING  
DATA MINING TECHNIQUES**

**BY:**

**GEDAM ESUBALEW MITIKU**

**A Thesis Submitted To Bahir Dar University, Bahir Dar Institute Of Technology, School Of Graduate Studies. In Partial Fulfillment Of The Requirements For The Degree Of  
Master of Science In The Information Technology In The Faculty Of Computing.**

**Advisor: Gebeyehu Belay (PhD)**

**©2023 Gedam Esubalew**

**NOVEMBER, 2023  
BAHIR DAR, ETHIOPIA**

## Declaration

I, the undersigned, declare that the thesis comprises my own work. In compliance with Internationally accepted practices, I have acknowledged and refereed all materials used in this Work. I understand that non-adherence to the principles of academic honesty and integrity, Misrepresentation/ fabrication of any idea/data/fact/source will constitute sufficient ground for disciplinary action by the University and can also evoke penal action from the sources which have not been properly cited or acknowledged?

**GedamEsubalew**

NameofthecandidateSignature

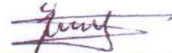


Date

November,2023

**APPROVAL SHEET**  
**BAHIR DAR UNIVERSITY**  
**BAHIR DAR INSTITUTE OF TECHNOLOGY**  
**SCHOOL OF GRADUATE STUDIES**  
**FACULTY OF COMPUTING**

**Approval of thesis for defense result**

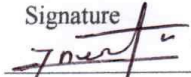
I hereby confirm that the changes required by the examiners have been carried out and incorporated in the final thesis. Name of Student: Gedam Esubalew Signature   
Date November 2023. As members of the board of examiners, we examined this thesis entitled "Developing a Predictive Model for Rape Crime Prevention Using Data Mining Techniques" We hereby certify that the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of Science in "**Information Technology**".

**Board of examiners**

Name of Advisor

Dr. Gebeyehu Belay (Ph.D.)

Signature



Date

27/02/2016 E.C

Name of External examiner

Dr. Tibebe Beshah (PhD)

Signature



Date

20/7/2023 G.C

Name of Internal Examiner

Mr. Belete Biazen

Signature



Date

28/02/2016 E.C

Name of Chairperson

Mr. Dagnachew Melesew

Signature



Date

27/02/2016 E.C

Name of Chair Holder

Dr. Abdurkerim Mohomed (PhD)

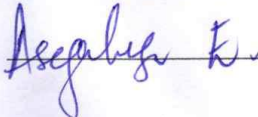
signature



Date

29/02/2016 E.C

Name of Faculty Dean



Signature



Date

29/02/16

iv



iv

## **ACKNOWLEDGEMENT**

First I would like to thank God for giving me the opportunity to follow my graduate study. I would also thank Dr. Gebeyehu Belay, my thesis advisor for his precious and constructive comments and encouragements all the way through my study. I have a loan from the deepest gratitude to my mother Wale Werkie for her valuable and strong advice throughout my life. I would also thank all the staffs of Addis Ababa Police commission for their good will for letting me use the data for this study and provide me with information about rape crime and related concepts. My deepest thanks go to my husband Getnet Birhanu who has supported me all the way to fulfill my dream, and I am highly thankful to my associates Tegegne Yemata, Anegaw Sisay and all my precious instructors for their unreserved support and encouragement throughout my study.

## **ABSTRACT**

Rape is serious crimes that have a troubling impact on victims in particular and the society in general. It is very dynamic and flexible. It is also sensitive for social and political crises. Therefore, developing effective rape crime prevention and mitigation methods to protect the people is indispensable. In relation to these, data mining techniques play a critical role to develop predictive models that can identify individuals who are at risk of being raped. The trend data modeling and analysis is essential to interfere and prevent the victim from various suspected rapes. In this thesis, a predictive model for rape prediction and prevention has developed using data mining techniques. The K-nearest neighbors (KNN) and decision tree (DT) algorithms have been used to develop the model. The data obtained from Addis Ababa Police Commission (AAPC) that has been collected from 11 sub-cities of Addis Ababa City. The data has organized and prepared in a table form of 10 columns and 4067 rows. The data has been pre-processed and trained so as to develop the predictive model. After the models have been developed, its performance has been evaluated using testing data set and accuracy. The KNN and DT models were able to predict the probability of a rape crime occurring with an accuracy of 93.5% and 88.1% respectively. Based on the result of model comparison score, KNN was more effective technique than DT. The results show that the system is able to accurately predict the rape crime in a given area. The models can be used to identify areas where prevention efforts should be focused and hence to design appropriate interventions mechanisms to prevent rape crime.

***Key Words: Rape Crime; Data Mining; K-nearest neighbors (KNN) and Decision Tree (DT).***

## CONTENTS

<b>ACKNOWLEDGEMENT</b> .....	<b>V</b>
<b>ABSTRACT</b> .....	<b>VI</b>
<b>CHAPTER ONE</b> .....	<b>1</b>
<b>INTRODUCTION</b> .....	<b>1</b>
<b>1.1. BACKGROUND</b> .....	<b>1</b>
<b>1.2. STATEMENT OF THE PROBLEM</b> .....	<b>3</b>
<b>1.3. OBJECTIVE</b> .....	<b>4</b>
<b>1.3.1. GENERAL OBJECTIVE</b> .....	<b>4</b>
<b>1.3.2. SPECIFIC OBJECTIVES</b> .....	<b>5</b>
<b>1.4. SCOPE AND LIMITATION OF THE STUDY</b> .....	<b>5</b>
<b>1.5. SIGNIFICANCE OF THE STUDY</b> .....	<b>5</b>
<b>1.6. THESIS ORGANIZATION</b> .....	<b>6</b>
<b>CHAPTER TWO</b> .....	<b>6</b>
<b>LITERATURE REVIEW</b> .....	<b>6</b>
<b>2.1. INTRODUCTION</b> .....	<b>6</b>
<b>2.2. DATA MINING FOR CRIME</b> .....	<b>6</b>
<b>2.2.1. K-NEAREST-NEIGHBOR</b> .....	<b>8</b>
<b>2.2.2. DECISION TREE</b> .....	<b>10</b>
<b>2.3. RAPE CRIME</b> .....	<b>13</b>
<b>2.3.1. TYPES OF RAPE</b> .....	<b>14</b>
<b>2.3.1.1. Diminished Capacity Rape</b> .....	<b>14</b>
<b>2.3.1.2. Age Related Rape</b> .....	<b>14</b>
<b>2.3.1.3. Incest</b> .....	<b>14</b>
<b>2.3.1.4. PARTNER RAPE</b> .....	<b>15</b>
<b>2.3.1.5. Aggravated Rape</b> .....	<b>15</b>
<b>2.4. THE RISK FACTORS OF RAPE</b> .....	<b>15</b>
<b>2.5. RELATED WORK</b> .....	<b>16</b>
<b>CHAPTER THREE</b> .....	<b>21</b>
<b>METHODOLOGY</b> .....	<b>21</b>
<b>3.1. INTRODUCTION</b> .....	<b>21</b>
<b>3.2. PROPOSED ARCHITECTURE</b> .....	<b>ERROR! BOOKMARK NOT DEFINED.</b>
<b>3.2.1. BUILDING A DATASET</b> .....	<b>23</b>



3.2.2.	DATA COLLECTION .....	24
3.2.3.	DATA SOURCES .....	25
3.2.4.	PRE-PROCESSING .....	26
3.2.5.	FEATURE EXTRACTION .....	27
3.2.6.	TRAIN TEST SPLIT .....	29
3.2.7.	APPLY CLASSIFICATION ALGORITHM .....	29
3.2.7.1.	K-Nearest Neighbors .....	29
3.2.7.2.	Decision Tree Algorithm .....	30
3.3.8.	EVALUATION METRICS AND MODEL PERFORMANCE .....	31
3.3.9.	IMPLEMENTATION TOOLS .....	32
<b>CHAPTER FOUR.....</b>		<b>33</b>
<b>RESULTS AND DISCUSSIONS.....</b>		<b>33</b>
4.1.	INTRODUCTION.....	33
4.2.	DATASET.....	33
4.2.1.	DATASET ANALYSIS AND FEATURE SELECTION .....	33
4.2.1.1.	Victim Age and Rape Crime .....	36
4.2.1.2.	Education Status and Rape Crime .....	37
4.2.1.3.	Rape Crime and Sub cities .....	38
4.2.1.4.	Rape Crime and Sex .....	39
4.2.1.5.	Family Condition and Rape Crime .....	40
4.2.1.6.	Rape Crime and Years .....	41
4.2.1.7.	Occupation and Rape Crime.....	42
4.2.1.8.	Offenders and Rape Crime .....	43
4.2.1.9.	Offenders Age and Rape Crime.....	44
4.3.	MODEL DEVELOPMENT .....	45
4.3.1.	K-NEAREST NEIGHBORS ALGORITHM IMPLEMENTATION .....	46
4.3.2.	DECISION TREE IMPLEMENTATION.....	51
4.4.	MODEL EVALUATION .....	52
4.5.	RESULT AND DISCUSSIONS .....	53
4.6.	USER ACCEPTANCE EVALUATION .....	55
<b>CHAPTER FIVE .....</b>		<b>56</b>
<b>CONCLUSION AND RECOMMENDATION .....</b>		<b>56</b>
5.1.	CONCLUSION .....	56
5.2.	RECOMMENDATION.....	56
<b>REFERENCES.....</b>		<b>58</b>
<b>APPENDIX.....</b>		<b>61</b>

## List of abbreviations

AAPC	-----	Addis Ababa Police Commission
ARFF	-----	Attribute-Relation File Format
DBMS	-----	Data Base Management System
DS	-----	Data Set
DT	-----	Decision Tree
EDA	-----	Exploratory Data Analysis
ETPC	-----	Ethiopia's Temporary Penal Code
GIS	-----	Geographic Information system
IG	-----	Information gain
KNN	-----	K nearest Neighbor
NB	-----	Naïve Baye
RF	-----	Random Forest

## List of Figures

Figure 2. 1	Euclidian Distance Formula .....	9
Figure 3. 1	Proposed architecture of rape crime data Analysis .....	23
Figure 3. 2	The collected dataset information .....	26
Figure 3. 3	Attribute correlation .....	28
Figure 4. 1	Feature Selection .....	34
Figure 4. 2	The correlation values of the dataset .....	35
Figure 4. 3	Victim age by number of crime.....	36
Figure 4. 4	Grade level by number of crime .....	37
Figure 4. 5	sub city by number of crime .....	38
Figure 4. 6	Sex by the number of crime.....	39
Figure 4. 7	Family condition and rape crime .....	41
Figure 4. 8	Number of victim per a year.....	42
Figure 4. 9	Rape crime and occupation .....	43
Figure 4. 10	Offenders and Rape Crime .....	44
Figure 4. 11	Offenders Age and Rape Crime .....	45
Figure 4. 12	Error rate.....	47
Figure 4. 13	Selection of K values.....	48
Figure 4. 14	Train Test Accuracy in KNN .....	49
Figure 4. 15	Confusion Matrix of KNN Model .....	50
Figure 4. 16	Visualizing the Decision Tree Confusion Matrix.....	51
Figure 4. 17	Decision tree structure .....	52
Figure 4. 18	Performance of Prediction Model .....	53

## List of tables

Table 3. 1 The collected dataset.....	24
Table 3. 2 General attributes of dataset .....	24
Table 3. 3 labeled dataset.....	27
Table 3. 4 Confusion Matrix.....	31
Table 3. 5 Description of the tools and python packages used during the implementation.....	32

# CHAPTER ONE

## INTRODUCTION

### 1.1. Background

Today, rape is a criminal act that receives a great deal of public attention. The true type of criminal offense has always existed or can be considered as a classic form of crime that would always evolve with human culture; it would always be present and expanding[1]. Rape is a criminal act that occurs not just in rural places where traditional values and customs still exist but also in relatively large cities with a more advanced culture and awareness of the law. Crime is a dangerous social problem that exists everywhere. Crime affects a country's reputation, economy and quality of life. As is generally known, rape is a common crime in today's socioeconomic development, particularly among the poor.

Rape is an important form of gender-based violence against women. While all forms of gender-based violence against women are serious, rape is especially hurtful and damaging, and can have long-lasting consequences. Few men are victims of rape [2], though this is much less common than among women. The issues that victims of rape encounter are quite complex. The issue arose not only as a result of the rape that occurred on the victims but also as a result of legal procedures brought against them. Victims of crime may be treated unfairly during the procedures, such as being bombarded with questions that cast doubt on their participation in the event and snapping during the trial.

Victims of rape do not report the incident to law enforcement officials for a diversity of reasons, including the victim's shame or the victim's fear of being killed if they report the incident to the police. This, of course, has an impact on the victims' mental growth and mental health, as well as the law enforcement process itself. Victim factors play a crucial role in overcoming or resolving these rape cases; this necessitates the victims' fortitude to report the crime to the police, as most victims of rape under threat would be devoted again by actors, causing fear and trauma in the victims. As a result of this complaint, the case can be opened and they can complete the inspection process, allowing victims to receive justice for what occurred to them. The victim can sue the prisoner for damages or compensation under positive law [3].

A methodical approach is important in identifying and analyzing trends and patterns in crime analysis and prediction. Data mining can be a great benefit in analyzing, visualizing, and predicting crime by using utilizing crime data sets. It is used to show crime trends, and significant hidden relationships between the crimes.

In industries with massive amounts of data, like law enforcement, information support systems have become essential for organizing and processing data for decision-making. Data mining technique is used to find patterns that are hidden in enormous databases. The process of automatic or semi-automated study and analysis of tremendous amounts of data to uncover relevant patterns and rules is known as data mining. So, police can review and handle massive amounts of data using data mining techniques to spot patterns in crime. Thus, this study would predict the prevention of rape crimes using data mining approaches.

Data mining is the process of extract and discovers hidden knowledge and information. Data mining also used to analyze and forecast the information and the item. Data mining used in different categories like prediction, clustering, relationship mining, and discovery models. Knowledge of the numerous techniques is required to choose the one that is best suited to the particular data mining issue[1]. Various modeling techniques and the most commonly used data mining techniques are decision trees, neural networks, genetic algorithms, nearest neighbor methods, and rule induction. However, decision tree and KNN modeling methodologies would be used in this study. A decision tree is a technique in which records are presented in a tree structure based on the values of their attributes [4].

A decision tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered. A model is a form of supervised learning, meaning that the model is trained and tested on a set of data that contains the desired categorization. Crime analysis and prevention is a methodological technique for identifying and studying patterns and trends in crime. Analysis and prediction of rape crimes helps the country's law enforcement agencies make more accurate decisions. B. Allocate more resources, such as police officers, to crime-prone areas.

This paper identified areas with high crime potential and predicts these areas. Criminal data analysts help police accelerate criminal investigations as more systems become automated. Various data mining techniques are used in this study. Various law enforcement agencies including Addis Ababa police use these results to guide their next rape crime analysis so that it supports crime prevention and to better understand the benefits of data mining in this area. Using crime analysis and prediction is a systematic way to find crime. Data mining techniques are used to identify areas with high crime potential and predict where crime will occur. The idea of data mining allows us to extract valuable previously undiscovered information from unstructured datasets.

## **1.2. Statement of the Problem**

Rape crime in cities like Addis Ababa has become a complicated social phenomenon in recent years. As a result, law enforcement agencies, such as the police, must understand the elements that contribute to rising crime rates. There is always a need for systematic rape crime prevention tactics and policies to regulate this social evil. One means of learning about rape crime victims and criminals is processing of rape crime records. In general, law enforcement authorities must seek to identify, forecast, respond to, and prevent female and male related offenses [5]. This is accomplished by studying previous criminal behavior patterns and mapping their predicted future occurrence through the identification of offenders' and victims' characteristics.

Currently, there are distinct challenges associated with crime prediction. Data analysis is difficult due to incomplete and inconsistent data, limitations in obtaining criminal data from law enforcement, and data accuracy that is highly dependent on the accuracy of the training set. There are also some technical problems with crime prediction and analysis.

Rape crime is one of the main problems governments all over the world have been confronted, in cities like Addis Ababa has become a complex social phenomenon and has serious multifaceted impact on the victim individuals.

Generally, rape crime has negative effects on the socioeconomic development of the society.

Some of the main socioeconomic impacts of rape crime are:

It reduces investment, reduces safety, disrupt order, creates confusion, generates stress, creates economic cost, physical injuries, fear and anger.

## **Causes of Rape Crime**

Some of the main socioeconomic factors that caused crime in general and rape crime in particular are: unemployment, lack of education, poverty, injustice, alcohol, use of drugs. The other main factor that contributes for rising rape crime rate in Addis Ababa city is lack of predictive data about rape crime prevention. Thus, understanding the factors that contribute to rising rape crime rates and developing systematic rape crime prevention policies and strategies is mandatory. Therefore, this paper developed predictive model for rape crime prevention using data mining techniques.

Data mining refers to obtaining or mining knowledge from enormous amounts of data, knowledge mining from data, knowledge extraction, data or pattern analysis, and data archaeology [6]. The practice of extracting interesting knowledge from enormous amounts of data housed in databases, data warehouses, or other information repositories is known as data mining. Data mining is used to aid police roles such as life protection, crime reduction, social order maintenance, and individual freedom and privacy protection by extracting hidden beneficial patterns from huge amounts of crime data stored in hard copy and maintained in an electronic format. This study aims to identify rape crime trends in the city, which are important for the rape crime prevention and control of crimes against women and men in the city using data mining technique. Therefore, this paper developed predictive model for preventing rape crime using data mining technique.

The research questions are formulated as follows:

- How to detect determinants of rape crime factors for predictive model development?
- How to use data mining techniques for rape crime analysis and model development?
- How to develop and evaluate a predictive model for rape crime?

### **1.3. Objective of the Study**

#### **1.3.1. General objective**

The general objective of this study is to develop a model and analysis of rape crime prevention in the cities using data mining techniques.



### **1.3.2. Specific objectives**

The following specific objectives were specified so as to achieve the above mentioned general objective.

- To detect the determinants of rape crime attributes.
- To evaluate and select data mining techniques which are appropriate to rape crime.
- To develop a predicting model for rape crime prevention.
- To train and test the models functionality.

### **1.4. Scope and Limitation of the study**

The paper focused on the analysis of rape crime in Addis Ababa city using data mining technique. KNN and DT data mining models were used for both crime analysis and prediction. The scope of this study is limited to Addis Ababa city on the prevention of rape crime. It was delimited to identify events associated with offenses against females as well as males with associated conditions. The study does include crime records that had not received final judgment by the court and also detects events linked to male-on-female crimes. The study used only classification mining techniques. It also focuses on only one type of crime i.e. rape crime.

Some of the main limitations of the study were the inconsistency of the rape crime data regarding the data of suspects, who often give false names, birth dates, places, and times to police officers and thus have multiple database entries, making it difficult for officers to determine a suspect's true identity and relate past incidents involving that person. The other limitation was that several criminal records are found in hard copy format. Data in soft copy are very scarce.

### **1.5. Significance of the Study**

Information retrieval and crime forecasting models are currently popular in our society. By forecasting future crimes using current crime dataset, it aims to lower the incidence of crime. Several organizations and police departments can use this result in preventing rape crimes. The main purpose of this study is to predict rape crimes that can occur in the future and support law enforcement agencies in preventing rape crimes before they occur. The capability to predict rape crimes based on time, location, and so on can help provide useful information to law enforcement agencies. In addition to this, it is used to extract violent crime patterns that relate to

criminal characteristics so that police officers can use these patterns in their day-to-day crime controlling activities.

## **1.6. Thesis Organization**

This thesis report consists of five chapters. The first chapter deals with the general overview of the study, including the background, statement of the problem, objectives, scope and limitations of the study, and significance of the study. The second chapter contains a literature review. The third chapter contains methodology. This chapter is about methodology (materials and methods). Chapter four reports the results and discussion. The results of the study are analyzed and interpreted. The last chapter presents conclusions and recommendations.

# **CHAPTER TWO**

## **LITERATURE REVIEW**

### **2.1. INTRODUCTION**

This chapter discusses how to clearly explain the overall research topic and the depth of the information to be presented; it often also explains the types of sources that will be used. It should contain an introduction, a body, and a conclusion in finding out the arrangements in rape crime prevention analysis. It helps to understand the existing research and debates relevant to a particular topic or area of study and to present that knowledge in the form of a written report.

### **2.2. Data mining for crime**

Data mining is the discovery of unexpected patterns and new rules hidden in large databases[7]. The practice of gathering information that they believe will either directly or indirectly benefit their businesses has become almost universal. Along with the ability to collect, a lot of

information is also being generated both inside and outside the organizational structure. There are many elements that contribute to the growth of large amounts of data.

The usage of bar codes, the automation of many businesses and other transactions, the improvements in data collecting tools, including scanned picture records on image platforms as well as satellite remote sensing systems, are all examples of this. Data mining is a method of problem-solving that identifies logical or mathematical explanations of patterns and regularities in a set of data, eventually of a complex nature. Three different sources are used to create data mining techniques: statistics, machine learning, and artificial intelligence.

Data mining refers to obtaining or mining knowledge from enormous amounts of data, as paper[6] explain simply. Knowledge mining from data, knowledge extraction, data or pattern analysis, and data archaeology are some other phrases that have a similar or slightly distinct meaning to data mining. Data mining is also a way to extract knowledge from usually large data sets; in other words, it is an approach to discovering hidden relationships among data by using artificial intelligence methods. The wide range of data mining applications has made it an important field of research. It is the knowledge discovery process used to collect and analyze a large dataset and summarize it with helpful information.

Data mining is essential to serve analytical purposes in various fields of science and plays an important role in human life as well as areas such as education, business, medicine, health and science[4]. Data mining is the extraction of hidden patterns and useful trends from large databases. It is a robust technology with great potential to enable organizations to focus on the most valuable information in their databases. Data mining extracts implicit and potentially useful information from vast amounts of data stored in databases by creating computer programs that automatically or semi-automatically search databases for meaningful patterns[8]. As databases have grown so large and new engines with search capabilities have been developed, the use of

data mining has greatly increased. Data mining is usually useful for large amounts of data. For this reason, most algorithms developed for data mining require large amounts of data to build and train models responsible for various data mining tasks such as classification, clustering, prediction, and association.

The DM's goals can be changed in light of the planned use of the framework. We expect unique strategies and procedures to reveal different types of examples. All services that collect information today collect information in gigabytes per hour. DMs can help reduce, explore, and refine inferences to find new examples and data within the information that transcends the barriers to human data processing.

### **2.2.1. K-nearest-neighbor**

K-nearest-neighbor is one of the most basic and simple classification methods and should be used as one of the first choices for classification studies when there is little or no prior knowledge about the distribution of the data[9]. KNN algorithms make very accurate predictions and can compete with the most accurate models. KNNs are useful for both classification and prediction. However, it is more commonly used for classification prediction. KNNs group data into consistent subsets and classify newly input data based on similarity to previously trained data. Inputs are assigned to classes that share their nearest neighbors. KNN is the best data mining and machine learning algorithm used in classification techniques. Data points are classified based on how adjacent data points are classified and similarity to previously stored data points[10]. Therefore, in this study, we use the KNN algorithm to develop a predictive model for preventing rape crime. Compared to other algorithms, K-Nearest Neighbor is easy to understand and use.

The algorithm's superior accuracy, tolerance to outliers, and lack of assumptions on the data are its advantages. Its memory requirements and expensive calculations are some of its drawbacks. This method allows you to enter both numeric and nominal values. Due to its ease of use and simplicity, the k-nearest neighbor (KNN) algorithm is one of the most commonly used classification algorithms. The default classifier for many domain specific problems is KNN.

The KNN algorithm is a popular nonparametric classifier used for classification and regression issues. The data collection containing the data points is divided into two subsets, training data and test data. A KNN examines all of the training data to calculate the distance between two points and anticipate the nature of the input data. The distance between two points is determined by the Euclidean distance, a metric of similarity. The Euclidean distance is the distance between two points. We can locate his two spots on the plane by calculating the length of the line segment that connects the two points and using the Pythagorean Theorem.

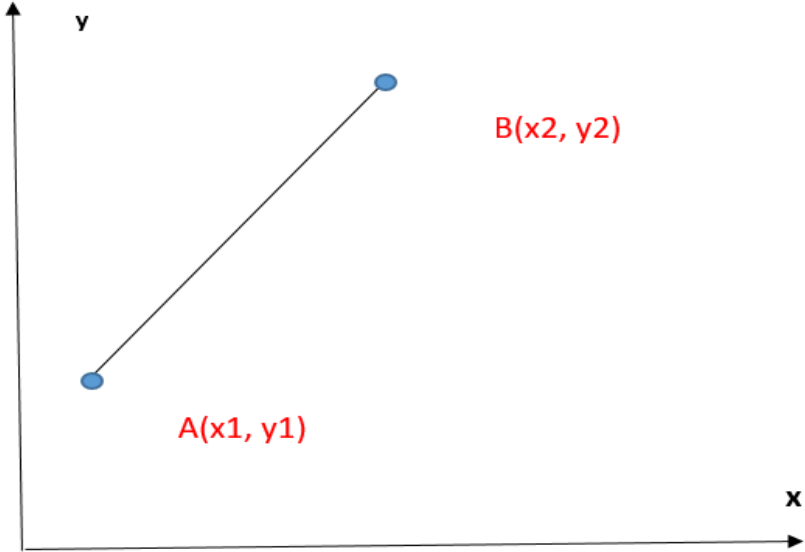


Figure 2.1: Euclidian Distance Formula

Euclidean distance: This is the most ordinarily utilized distance measure, and it is restricted to genuine esteemed vectors. Utilizing the underneath equation, it estimates a straight line between the inquiry point and the other point being estimated.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \text{ -----Eq (1)}$$

Once the distance between the input data and the training data is computed using the above algorithm, the k points closest to the input data are selected, and the majority class of the selected neighbors is the predicted class for the new data.

There are a training data set and a test data set for the k-Nearest Neighbors (KNN) algorithm. Each instance of training data has multiple features and labels. The overall goal of creating this algorithm is to label data that has not yet been supplied with new data. The procedure begins by initializing the integer value k. The next step is to compute the Euclidean distance between each set of training data given to the algorithm and the input data. The distances between each training set and the input data are then sorted, after which the k nearest neighbors of the input data are selected, where k is practical. After the adjacent data are selected, the input data are assigned labels representing the majority of the adjacent data.

### **2.2.2. Decision Tree**

Decision trees are used for both classification and prediction. For classification purposes, we can train a function using intervals divided according to a person's attributes. Decision support tool known as a decision tree that uses a tree-like model to represent the occurrence of options and their attributes, such as utilities, resource costs, and random event outcomes. A technique for demonstrating an algorithm that uses only conditional control statements is to use this method. The result of a decision tree is a tree-like structure that represents the series of decisions made at each step. These decisions are considered guidelines for the classification process. [11]. A decision tree model contains rules for predicting a target variable. Decision tree classification algorithms provide easy-to-understand explanatory data. It is intuitive that classifying large datasets improves the accuracy of classification models.

A decision tree is an easy-to-understand tree-structured classification model that is efficiently derived from data. It is one of the most popular algorithms for learning uniquely developed discriminative models in statistics. Decision trees are very easy to express and understand, making them more powerful than other algorithms used for the same problem[12]. Based on this, this study uses a decision tree algorithm to develop a predictive model for rape crime prevention. The most common approach to determining the class of a particular feature is a decision tree representing the design process. Each node in the decision tree is either a leaf node containing

the name of the class (aka response node) or a non-leaf node containing the attribute test and a branch to another decision tree (aka decision node).

Leaf nodes in decision tree diagrams are typically ovals, while decision nodes are typically rectangles with arrows that emphasize the relationships between related nodes. The development of ID3 (Iterative Dichotomizes 3) is C4.5, a divide-and-conquer strategy for guiding decision trees. It is widely used by several commercial software programs and is growing in popularity. This study uses C4.5, which was chosen because it contains additional properties. The fundamental strength of machine learning models comes from their underlying learning methods. The underlying approach for this decision tree implementation is the ID3 algorithm. The ID3 algorithm splits the data set according to attributes and decides when to finish splitting.

Decision trees are one of the simplest and most popular machine learning algorithms, and are primarily used for making predictions on categorical data. Entropy and information gain were the primary measures used in determining the relevance of decisions when building decision tree models. The idea of a decision tree is to split a data set into smaller data sets based on descriptive features until you arrive at a small enough set containing data points that match your labels. Each feature in the data set becomes the root node [parent node] and the leaf nodes [child nodes] represent the results.

Entropy is a measure of impurity, disorder, or uncertainty in a set of samples. Entropy controls how the decision tree decides to split the data. Entropy is a metric used in information theory to measure how pure or uncertain a set of observations is. Controls how the decision tree determines how to split the data. Information theory techniques are used to compute the information gain, the difference in the amount of information before and after splitting, to split the data set according to the optimal attributes.

**Entropy(x) =  $\sum_{k=1}^k p_i \log_2 p_i$  - - - - - - - - - -Eq (2)**

Where p is the probability occurrence of class i. Entropy is measured between 0 and 1 that depending on the number of classes in the dataset.

The amount of knowledge that gives a function about a class is measured by information gain (IG). This demonstrates the importance of certain feature vector properties. It used to determine

the order of attributes within the nodes of the decision tree. The amount of knowledge a function conveys about a class is measured in terms of information gain. Information-gathering strategies can be used to determine the layout of attributes within the nodes of the decision tree. A primary node is called a parent node, and a child node is called a child node. The information gathered affects both node splitting and decision tree splitting quality.

Information Gain = entropy (parent) – [average entropy (children)]

$$IG = E(p) - aE(C) \text{ ----- Eq(3)}$$

Where IG is information gain, E (p) entropy of parent and aE(C) is average children entropy.

The decision tree algorithm works like this:

1. To split the dataset, use the Gini index or cross-entropy to select the best features.
2. Create a feature selection node to split the dataset into more manageable pieces.
3. Once all columns have been committed to the same feature value, or there are no more features or occurrences, the operation is repeated for each child to start building the tree.

Gini contamination and information gain are used for subgroup selection.

The higher the entropy has more diverse the dataset. After classifying data instances belonging to the same class at the same leaf node using the decision tree technique, the split with the largest information gain was selected as the best feature to split.

The best features to share are selected using an ID3 approach that is also used for sharing itself. Shannon entropy, also called entropy, is used to determine the amount of information in the data stack. Before running an algorithm, prepare the dataset for processing by collecting input data and training the algorithm on the specified dataset. Use the information gain calculation to select the appropriate features to split. To get more information, split the dataset into groups according to the most favorable properties, make sure all the data in each group belong to the same class, and stop splitting if necessary. This is not the case until you decide to use new features to split your dataset, split it further, and create new branches as new classes become available. As a result, this process continues until all terminal nodes in the tree have components of the same class.



### **2.3. Rape Crime**

Crime is a well-known social problem that affects societies' living standards and economic growth in many ways. Direct or indirect physical or psychological cruelty to women is what is meant by the term crimes against women [13]. A "crime against women" is defined as a crime committed specifically against women, where women are the only. It is an offence that may be prosecuted by the state, especially if it is a serious violation of the law and a serious crime, especially if it violates good morals and is an attempt to combat crime. Internationally, crime rates are higher in cities than in rural areas, with the rate generally increasing with city size[14].

Crime is social injustice. This is a crime that violates state law and is highly frowned upon by society. A crime is defined as an act or omission that violates the law and is subject to fines or imprisonment. Criminal behavior is considered anti-social behavior. Crime can be legal or illegal. An unlawful and punishable offence is a violation of the administrative or legal system of a state or committing an offence harmful to oneself or a third party within the meaning of criminal law. All forms of self-defense are permissible legal and non-criminal offences. No one is born a criminal. Instead, criminal behavior and criminal intent are the result of many social, economic, biological, and psychological factors[15].

Rape crime is a complex behavior that can be exacerbated by many factors, including personal development and family history, personality traits, and environmental and cultural influences. Rape is a very serious form of violence. This is an important form of gender-based violence against women. All forms of gender-based violence against women are serious, but rape is particularly traumatic, harmful and can have long-lasting consequences[16]. Rape is a high-profile crime. Various research results show that real-life types of crime have always existed or can be viewed as classic forms of crime that have always followed the evolution of human culture. It probably won't be much different than it used to be, but it will always be there and expand.

Rape is any unlawful sexual act, including sexual intercourse through violence or the threat of violence, or with someone who is unable to give legal consent because they are underage, mentally ill, mentally retarded, intoxicated, or unconscious. In many agencies, the crime of rape is included in the crime of sexual assault. Rape has long been viewed as caused by unbridled

sexual desire and claims of extreme power over the victim. Many are unable to effectively deal with the aftermath of their rape, but continue to fight in silence. Rape is defined as violence, threat of physical harm, or oral or anal penetration without consent obtained when the victim cannot consent[17]. Rape can also be defined as the use of physical force, fear, or deception to obtain sexual relations with a woman against her will and without her consent.

### **2.3.1. Types of Rape**

Forms of rape can be identified based on the person who committed the rape, the rape victim, and specific acts associated with rape. Some types of rape are considered much more serious than others.

#### **2.3.1.1. Diminished Capacity Rape**

Diminished rape occurs when a person is forced to engage in sexual intercourse with another person who does not consent to sexual activity. An incapacitated person is unable to consent to sexual activity because of their limited physical or mental capacity. It also occurs when people are unable to consent to sexual activity because of alcoholism.

#### **2.3.1.2. Age Related Rape**

Another form of rape is age related. This type of rape is often called statutory rape. In this case, sexual activity with a person under the minimum age is always considered illegal. Under Ethiopian law, the minimum age for sexual consent is 18 for her. According to Ethiopia's Temporary Penal Code (ETPC), minors under the age of 18 are considered incapable of consenting to sexual intercourse. Crime and Violence Scales found that women aged 16 to 19 were the most likely to be rape victims, and rape perpetrators were likely to be in this age group as well[18].

#### **2.3.1.3. Incest**

Incest is a type of rape determined by the relationship between the two. If both parties involved in the sexual activity are closely related (that is, if they are family members).

- Parents and children
- Uncles and nieces or nephews
- Aunts and nieces or nephews

#### **2.3.1.4. Partner Rape**

Partner rape also known as spousal rape, is a type of rape involving a person's partner or previous partner. There are three types of partner rape:

- Beating rape – involving both physical and sexual violence
- Force-only rape – involving the imposition of power and control over another
- Aggressive rape – involve suffering and headstrong sexual acts.

#### **2.3.1.5. Aggravated Rape**

Aggravated rape is a type of rape defined in the law. Aggravated rape involves:

- Forced sex acts by the threat of death or serious bodily injury
- Forced sex acts involving an unconscious or drugged victim
- Sex acts with children under the age of 18

### **2.4. The Risk Factors of Rape**

Various social factors have been associated with an increased risk of sexual assault as adults. Most prominent are traditional norms regarding gender roles, the existence of male sexual rights, weak legal sanctions against sexual assault, and social norms in favor of sexual assault. Below are the most common risk factors for rape crimes[19]

**Age:** One of the main risk factors for rape is age. The Crime and Violence scale found that women aged 16 to her 19 were the most likely to be rape victims, and the perpetrators were also more likely to be in the same age group. This age group is most likely to report rape to the police[18]. Socioeconomic factors were also risk factors. Women from low-income households were at higher risk of rape than women from high-income households[20]. Low socioeconomic

status has also been identified as a risk factor for harm, women from Other Violent Crimes Households.

**Alcohol:** Alcohol consumption can increase the risk of sexual assault in many ways. When potential aggressors consume alcohol, they may experience more aggressive behavior and be unable to accurately interpret the sexual interests of others. Another controversial risk factor is pre-crime alcohol consumption[21]. The criminal justice response to rape is important because it has lasting consequences for the victim, not only in terms of trauma, but also depression and increased likelihood of future behavioral problems as a coping mechanism. Many researchers have conceptualized the importance and application areas of rape crime prevention at various times. Various research results show that the actual forms of crime have always existed or can be regarded as classic forms of crime that have always followed the development of human culture. This study uses KNN and classification techniques to analyze rape crime prevention.

## **2.5.Related Work**

This part provides a methodical analysis of the fundamental literature in the field of rape crime identification. Additionally, this subject will be revealed in order to identify the knowledge gap as well as properly explain the general methodologies, techniques, and findings of previous investigations.

Khan et al [22]the researcher conducted the predictive model on rape crime prediction using data mining techniques. The researcher uses the United States rape crime dataset. The dataset has different attributes like the victim, the offender, and the circumstances of the crime. The study used a variety of data mining techniques to develop the predictive model. The most effective technique was logistic regression which was able to predict the probability of a rape crime occurring with an accuracy of 80%. This paper has a weakness a study was conducted on a dataset of rape crime data from the United States. It is not clear whether the findings of the study would be generalization to other countries. And also, the study did not evaluate the effectiveness of the interventions that were developed based on the findings of the predictive model. It is not clear whether the interventions were effective in reducing the risk of rape crime.

Saltos and cocea[1]The researcher extracts knowledge and analyzes the information for a better understanding of the crime and to possibly stop future crime. The researcher uses three algorithms from different categories of approaches: instance-based learning, regression, and decision trees.

Sathyadevan et al. [10] the researcher analysis and predict the crime by using data mining techniques like Naïve Bayes, Apriori algorithm, decision tree. In this paper, the researcher predicts regions with a strong probability of crime and can identify crime-prone regions. The researcher focuses on some crime factors like criminal background of offender, political enmity. In this paper, the researchers investigated the precision of categorization and prediction based on various test sets. The Bayes theorem, which demonstrated more than 90 % accuracy, is used to perform classification.

Zaharan et al.[11] The researchers detect and predict the crime by using data mining techniques like NB, KNN, decision tree, random forest, linear regression, logistic regression, SVM. In this paper, the researchers analyze and discuss the various variables that impact criminal activity as well as the techniques used to predict criminal activity and their consequences. In this paper, the researcher concludes that random forest classification performs with higher accuracy than other methods. The accuracy got 87% on the Los Angeles dataset, 90% on the Egypt dataset, 91% on the Chicago dataset, and 81.7% on the United States dataset.

Hossain et al. [19] The researcher develop decision tree, random forest and KNN model for the prediction of crime by using spatial temporal data. The researcher also analyzed the collected data for the increasing of accuracy. The researcher has got 68.03% with unbalanced dataset.

Khan et al. [23] The researcher developed predictive model by using classification techniques and by using Sanfransisco crime dataset. By examining and contrasting three well-known prediction classification algorithms like Naive Bayes, Random Forest, and Gradient Boosting Decision Tree the researcher develops a crime prediction model. The model that the researcher created was based on an analysis of the top ten offenses, which accounted for 90% of the instances. Using a crime dataset, the researcher uses exploratory data analysis (EDA) to identify patterns and recognize patterns in crime. The proposed model is also assessed for

precision and recall matrices. The accuracy of Naive Bayes, Random Forest, and Gradient Boosting Decision Tree approaches are 65.82%, 63.43%, and 98.5%, respectively.

Yerpude[24]The researcher use data mining techniques for crime data that helps to predict features that affect the high crime rate. Research uses decision trees, Naive Bayes, and regression data mining approaches to make predictions about the variables that cause crime in a region or community using previously gathered data. The Crimes Record Bureau and Police Department can take the appropriate steps to reduce the likelihood that a crime would happen based on the rankings of the attributes. The researcher developed a Random Forest Classifier that gives the most balanced result of 83.39% and 86.54% with respect to accuracy and F1 score respectively.

Kim et al. [25]For the crime analysis and for the prediction of crime the researcher use 15-year dataset. The researchers investigate machine learning algorithms like K-nearest neighbor and enhanced decision trees for predicting crime ranges is from 39 % to 44 %. The researchers have discussed several problems with crime prevention and analysis. And also have proposed crime prediction algorithms. In this paper, several techniques and algorithms had significantly variable accuracy, complexity, and train times. The algorithm and the dataset can be tuned for particular applications to increase prediction accuracy. Despite the fact this model's prediction accuracy is poor. Researchers were finding some feature and the relationship of the given dataset by using data mining techniques. The paper discussion was data mining techniques for the analysis of the rape crime.

Castro and Hernandez [26] Researchers have developed a predictive model for assessing child conflict and child risk using data mining techniques for the Philippines. Researchers use data mining techniques to extract hidden features from policies. The purpose of thesis is to develop data mining models, such as Decision Trees, Naive Bayes, General Linear Models, and Logistic Regression, using datasets provided by Social Welfare. Also, check the performance of the predictive model. The researcher got 92% of accuracy and 7.35% of classification error in naive bays algorism. Research has found that children aged 15 to 17 also committed violent crimes and were victims of many abuses by the ages of 12 to 17.

Tayal et al. [27] The researcher develop data mining algorithm for the prediction and for the identification of crime. Researchers propose models by dividing the task into different modules, such as data extraction, data preprocessing, clustering, rendering Google Maps, classification, and implementation. Researchers used the K-Means data mining technique to detect and predict crime with a 90.9% performance.

Table 2.1: Summary of related work

<b>Authors name</b>	<b>Method</b>	<b>Problem</b>	<b>Accuracy</b>	<b>Finding</b>	<b>Gap</b>
Saltos and cocea, 2017	instance-based learning, regression, and decision trees	Extract knowledge, and analyze information	-	Clearly specify the data set and has large number of dataset	Doesn't specify the accuracy of the dataset
Sathyadevan et al, 2014	NB, APrior and DT	Analyze and predict the crime. Identify the most crime prone region	90%	investigated the precision of categorization and prediction	Does not specify the dataset clearly. Use small number of attributes

Zaharan et al, 2021	NB, KNN, DT, RF, SVM linear regression, and logistic regression,	Detect and predict crime. analyze and discusses the various variables that impact criminal activity	91%	High accuracy	doesn't specify the dataset clearly and use different unclear dataset
Yerpude, 2020	DT, RF, NB, and regression	Predict the attributes that affect the crime.	83.39%	Predict the attribute correctly.	Lowest accuracy. Need dataset normalization and hyper parameter tuning data
Castro and Hernandez, 2019	DT, NB, general LM, logistic regression	Develop predictive model, to extract hidden features	92.0%	The researcher addresses the problems and creates solutions.	doesn't state the crime and also does not clearly specify the dataset
Tayal et al, 2015	K-means	Prediction and Identification of crime,	90.9%	The researcher addresses his task	But doesn't specify the number of data, attributes,
Hossain et al, 2020	RF,DT and KNN	prediction of crime by using spatial temporal data	68.03%	analyzed the collected data for the increasing of accuracy	Use unbalanced dataset
Kim et al, 2018	KNN and DT	Data analysis and crime prediction using different algorithms	44%	use large dataset and prediction accuracy is poor	Does not describe the dataset clearly and use small features



The existing paper is a valuable contribution to the literature on rape crime prevention and prediction. The existing paper provides evidence that data mining techniques can be used to develop effective predictive models for rape crime prevention based on the context. The predictive models on the existing paper can be used to identify areas where prevent and predict the rape crime. On the existing paper the models can also be used to develop interventions to prevent rape crime. The existing paper has different limitation; the study was conducted on a dataset of rape cases from different region and country. It is not clear whether the findings of the study would be generalization to other cities or countries. And also, the study did not evaluate the effectiveness of the interventions that were developed based on the findings of the model. It is not clear whether the interventions were effective in reducing the risk of rape. This thesis used large dataset and different attributes, normalize the data set, determine the parameter optimization by using K-fold cross-validation, analyze the confusion matrix to identify potential areas for improvement in the models, the findings of the study would be described clearly and evaluate the performance of the model by using test set and accuracy.

## **CHAPTER THREE**

### **METHODOLOGY**

#### **3.1. Introduction**

To accomplish research goals and provide a solution to the research issue, many tasks involving such as collecting data, preprocessing, procedure of work describing, solve a research problems systematically and studying how research is done scientifically, non-scientifically and social problems, predicting phenomenon and algorithms. The algorithms utilized to perform the study on rape crime prevention is described and justified in the section. To create a model and predict rape crime prevention using data mining, the collected data needs to be preprocessed to remove unrelated information. Feature extraction and classification using different methods are also

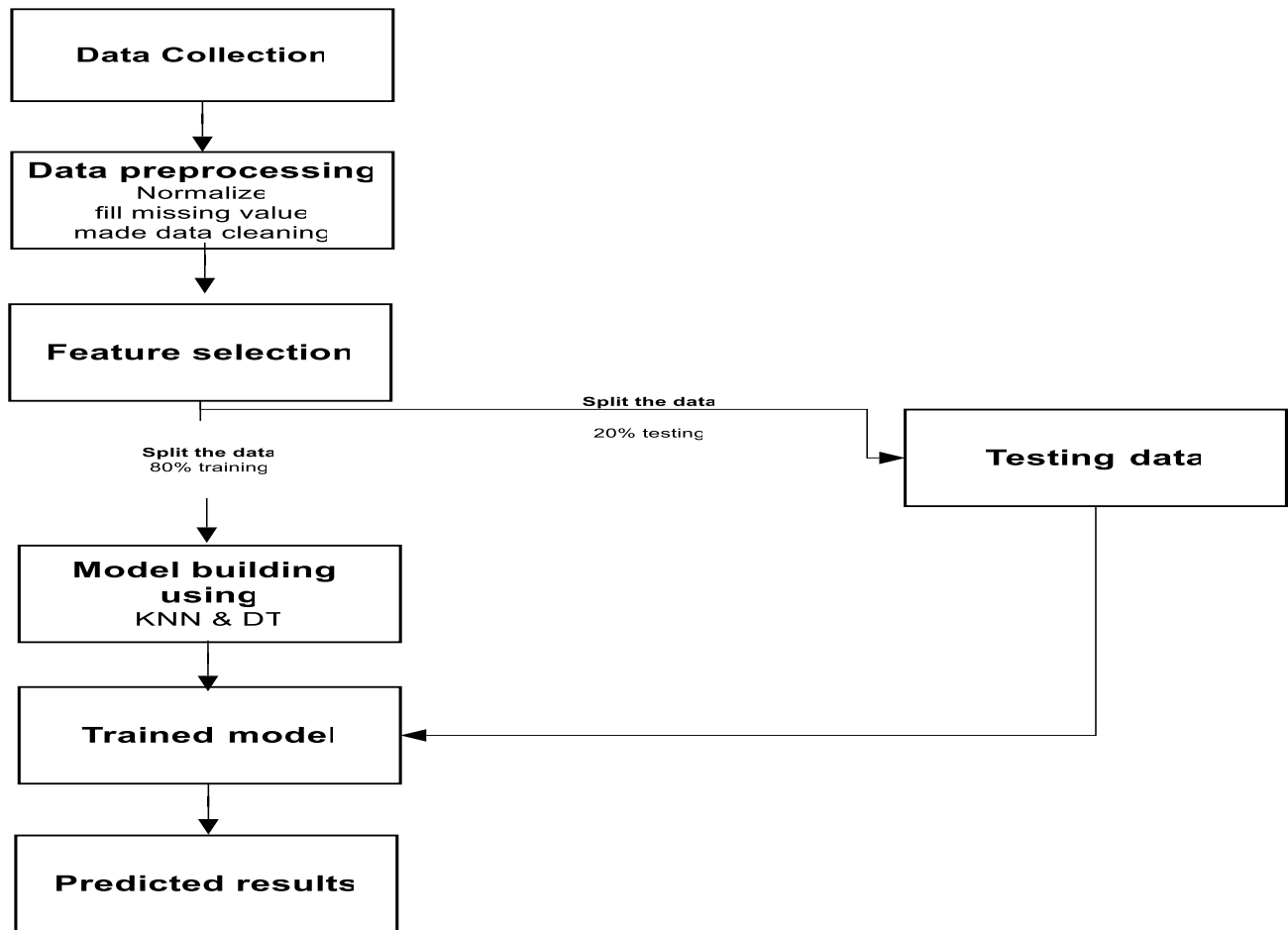
required. This section discusses the data mining and also how data was collected for the prediction of rape crime. It also discus about preprocess and analysis of parameters.

### **3.2. Steps of rape crime data analysis**

The proposed architecture aims to develop a predictive model for rape crime prevention in Addis Ababa city is depicted in the following Figure 3.1. Classification technique and different algorithms like KNN and DT were used to predict and build the model. The original dataset has trained and processed differently for each algorithm. This is achieved by implementing a Python script. Then the method was developed and a test model was generated using the data. Finally the model was tested using test scores to build the model and predict the number of specific crimes that occur in the city in a given year.

Understand the information contained in criminal records that perform processing tasks to detect crime trends and crime city so that can predict hot spots of upcoming criminal activity and reduce crime rates. Collected law enforcement data comes in a variety of formats and pre-processing data to improve the quality of the data set and efficiently collect the exact results you need. Extraction is the process of transforming text data in to real vectors using various feature extraction algorithms. Data classification is use to organize, structure and extract feature-extracted datasets.

The test data set was used to demonstrate how the trained model performed on all defined test data and to check how well the model made predictions. Models are designed to perform operations when feature selection, classification, analysis, prediction, and evaluation are performed and graphs using visualization tools visualize the results.



**Figure 3.1: steps of rape crime data Analysis**

### **3.2.1. Building a dataset**

The purpose of this study is to develop a predictive model for rape crime prevention using data mining. Therefore, creating a new rape crime prevention dataset is important. This new record is necessary because there is no record available or descriptive for this purpose. The process of building the rape crime prevention dataset involves collecting credible Addis Ababa Police Commission data, preparing and filtering the collected data into a file dataset, and interpreting the dataset.

### 3.2.2. Data collection

In this paper, the prediction of rape crimes has examined using multiple datasets from Addis Ababa police commission document records. The dataset was collected by a police commission that documented rape crimes in the city of Addis Ababa. The collected data is entered into a database for later processing. Because the data collected is unstructured. Data collection is the best choice because criminal data is unstructured and the number of fields, content, and size of documents can vary from document to document. A real-time rape crime dataset from Addis Ababa Police Commission from 2006 to 2014 E.C is used and the dataset size is 4067.

**Table 3.1: The collected dataset**

Date	Crime_type	victim_age	Sex	Grade_level	Work	Family_condition	Offender	ofendersage	Subcity
1/2/2006	Rape	7 f		2	student	family	n/know	65	Kirkos
1/3/2006	Rape	16 f		9	student	family	n/know	69	Kirkos
1/2/2006	Rape	4 f	kg		student	family	student	17	Lideta
1/10/2006	Rape	4 f	kg		student	family	student	18	Kolfe
1/11/2006	Rape	11 m		5	student	family	n/know	27	Lideta
1/4/2006	Rape	4 f	kg		student	family	n/know	22	N/silk
1/5/2006	Rape	23 f	degree		privat	self	n/know	55	Lideta
1/6/2006	Rape	18 f		12	privat	family	n/know	16	Akaki
1/5/2006	Rape	13 f		5	student	family	n/know	43	Akaki
1/6/2006	Rape	8 f		1	student	family	partner	24	A/ketema
1/13/2006	Rape	16 f		2	student	family	n/know	35	Kirkos
1/12/2006	Rape	14 f		8	student	family	nighbor	34	Yeka
1/14/2006	Rape	6 m	kg		student	family	student	25	Arada
1/16/2006	Rape	6 f	kg		student	family	nighbor	24	Arada
1/17/2006	Rape	17 f		8	student	family	family	29	Yeka
1/15/2006	Rape	15 f	No gl		No work	street child	n/know	45	Arada
1/17/2006	Rape	12 f	No gl		No work	street child	n/know	26	Yeka
1/6/2006	Rape	17 f		9	student	family	n/know	49	A/ketema
1/5/2006	Rape	7 m		1	student	family	n/know	25	A/ketema
1/5/2006	Rape	20 f		6	privat	relative	student	52	Akaki
1/15/2006	Rape	23 f		8	privat	self	nighbor	46	Akaki
1/11/2006	Rape	20 f		12	privat	parasite	family	37	Gulele
1/14/2006	Rape	23 f		10	privat	self	n/know	43	A/ketema
1/23/2009	Rape	22 f	diploma		privat	parasite	n/know	43	Gulele
1/22/2006	Rape	28 f	degree		privat	self	nighbor	28	Kirkos
1/18/2006	Rape	22 f	No gl		home servant	employer	employer	55	Kirkos
1/9/2006	Rape	27 f	degree		privat	self	n/know	63	Bole

The dataset has 4067 number of rows with 9 numbers of columns. The attributes are victim age, grade level, and work position, family condition of the crime, offender, offender age, and region of the crime. The first step is to gather data on rape crimes, such as the location, time, and date

of the crime, as well as any other relevant details, such as the age and gender of the victim, the relationship between the victim and the perpetrator, and any other contextual information that may be useful.

**Table 3.2: General attributes of dataset**

No	Attribute	Description
1	Date of crime occurred	When the incident occurred
2	Victim age	Age of victims
3	Gender	Gender who victims by offenders
4	Grade level	Grade level of victims
5	Offender	Who offended on victims
6	Work	Work of victims
7	Family condition	Victim's living condition
8	Offender's age	Age of offender when the incidence is occurred
9	Sub city	Sub city where the incident occurred

### 3.2.3. Data sources

The data source for building rape crime dataset has collected from Addis Ababa police Commission unpublished crime records.

- The data include both the victims as well as the criminal's profile.
- It contains the time and the place where the crime occurred.
- It also contains victim's age, grade level, victims family condition and offender's age

### 3.2.4. Pre-processing

The results of classification problems can be influenced by the criteria of the data set. Missing values affect the results. Therefore, the missing parameters of the data set must be managed first. Misplaced values can be treated in various ways, such as ignoring, replacing with an arbitrary number, replacing with the maximum value of the feature, or replacing with the mean of the feature. Missing values from the numerical data of this study are corrected by inserting the mean of the characteristics. This method, which gives better results than deleting rows and columns, can compensate for the loss of data. When counting 1s and 0s in Boolean data, missing values are replaced with the highest number of 1s or 0s that can be counted.

The paper analyzed and predicted the crime all over Addis Ababa that includes eleven sub cities namely Yeka, Bole, Arada, Gulele, Lemikura, Nifassilik, Kikors, Kolfe, Lideta, Akaki and Addis Ketema.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4067 entries, 0 to 4066
Data columns (total 10 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Date                  4067 non-null   object
1   Crime_type            4067 non-null   object
2   victim_age            4067 non-null   object
3   Sex                   4067 non-null   object
4   Grade_level           4067 non-null   object
5   Work                  4067 non-null   object
6   Family_condition      4067 non-null   object
7   Offender              4067 non-null   object
8   Offenders_age         4067 non-null   object
9   Subcity               4067 non-null   object
dtypes: object(10)
memory usage: 317.9+ KB
```

**Table 3.2: The collected dataset information**

Once the data has been collected, it needs to be preprocessed to ensure that it is in a format that can be analyzed by data mining techniques. This may involve cleaning the data, removing duplicates or outliers, and converting the data in to a suitable format for analysis. It is apparent from the value estimates that a significant proportion of the values in the date column are missing or null. There are many ways to deal with missing data, but the simplest is to remove partial rows. Rows containing null dates were eliminated. Moreover, the dataset was

filtered in order to include the columns which are most interested and to reduce the amount of irrelevant data.

Labeling a dataset means converting column data to numeric values and values to categorical values using label encoding techniques. This process is used to convert the text to numeric text for the normalization process using the label encoder () function. Where N is the number of class features. Convert the string into the numeric data type. The data types of the data set should be changed into number and integers.

Table 3.3:Labeled dataset

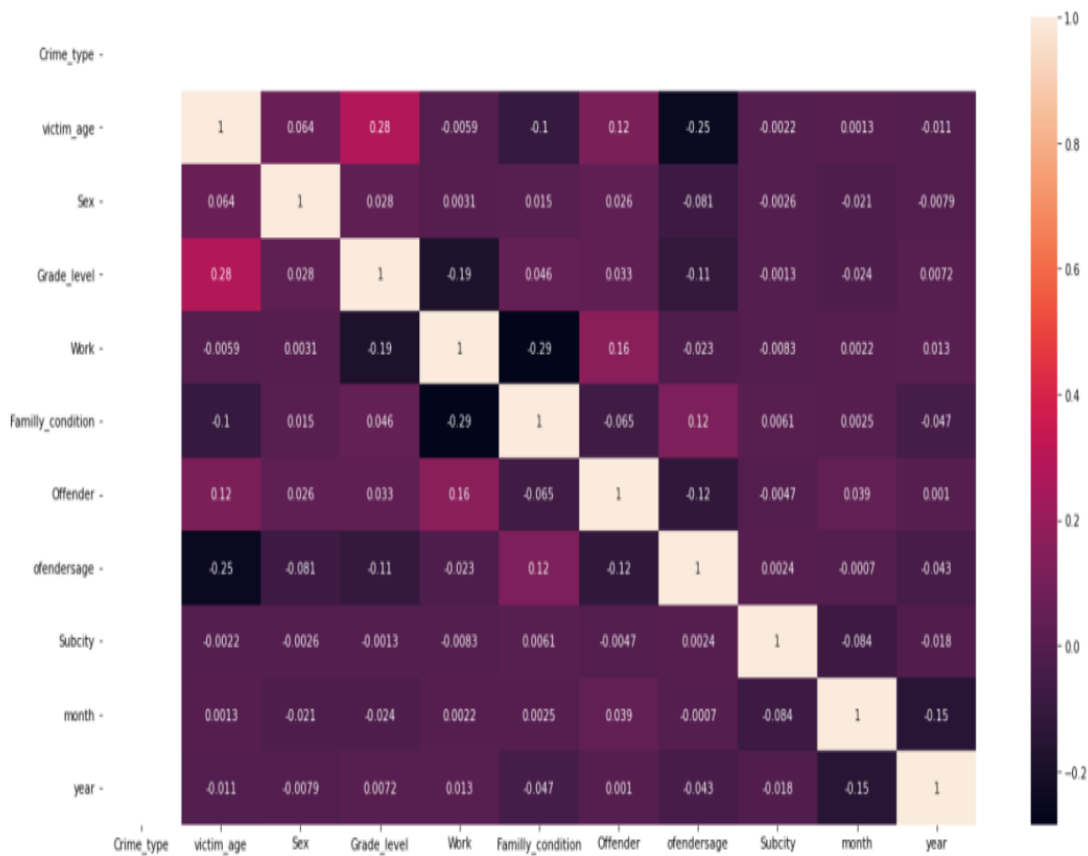
	Date	Crime_type	victim_age	Sex	Grade_level	Work	Family_condition	Offender	ofendersage	Subcity
0	2006-01-02	0	6	0	4	9	4	6	53	5
1	2006-01-03	0	15	0	11	9	4	6	55	5
2	2006-01-02	0	3	0	14	9	4	13	7	7
3	2006-01-10	0	3	0	14	9	4	13	8	6
4	2006-01-11	0	10	1	7	9	4	6	17	7
...	...	...	...	...	...	...	...	...	...	...
4062	2014-01-13	0	16	0	8	9	4	6	12	0
4063	2014-02-13	0	12	1	8	9	4	6	19	9
4064	2014-02-13	0	14	1	10	9	4	13	16	9
4065	2014-04-13	0	11	0	6	9	4	7	15	9
4066	2014-05-13	0	9	0	5	9	4	7	21	9

4067 rows x 10 columns

### 3.2.5. Feature Extraction

Reducing inputs for analysis and processing, or excluding the most important data, is called feature selection. Furthermore, selecting a subset of relevant properties to use in building a model is called feature selection; attribute selection, variable selection, or variable subset selection. The resource cost of representing huge datasets is reduced by feature extraction.

Analyzing a large number of variables generally requires a lot of computational power. Classification algorithms can over fit the training samples and perform poorly on new cases. In this thesis, a decision tree approach was used to extract features. Gini impurities used to automatically determine if a feature is suitable for separating objects representing different classes. Pearson correlation was also used for the extraction of feature. If the value is close to negative, the correlation of the attribute is weak. Otherwise, if the value is close to positive the attributes strongly correlated.



**Figure 3.3:Attribute correlation**

The two variables change in the same direction when there is a positive correlation. In a neutral correlation, no correlation exists between the changes in the variables. Variables change in the opposite way when there is a negative association. A positive correlation implies that as one variable rises, the other rises as well. A negative correlation, on the other hand, suggests that as one measure rises, the other falls.



### **3.2.6. Train Test Split**

After preprocessing, split the data set into a training data set and a test data set based on the specified split ratio. In this study, test used as the test data set and divided the data set into training and test sets according to the training rate. The preprocessed data required to split the data set into training and test data sets at 80% and 20% proportions, respectively.

The division of the data into training and testing sets is a crucial step in developing a predictive model for rape crime prevention using data mining techniques such as KNN and DT. The most common split for the data is 80% for training and 20% for testing, but other splits such as 70% for training and 30% for testing can also be used. The reason for using an 80/20 split is that it provides a good balance between having enough data to train the model and having enough data to test the model's performance. With an 80/20 split, we have a large enough training set to ensure that the model is adequately trained while still having a sufficient testing set to evaluate the model's performance. However, the choice of the split ratio depends on the size and quality of the dataset as well as the complexity of the model being developed. In some cases, a 70/30 split may be more appropriate if the dataset is small or if the model is relatively simple.

### **3.2.7. Apply classification algorithm**

For the prediction and analysis of rape crime, decision tree and KNN data mining techniques have been used. The extracted data can be analyzed by data mining techniques using classification, clustering, association rule mining, and visualization. From this, the study can be analyzed by classification techniques. Classification is a technique used in data mining to analyze collected data. This article uses two classification algorithms to predict rape crime prevention: the K Nearest Neighbors algorithm and the decision tree algorithm.

#### **3.2.7.1. K-Nearest Neighbors**

Both regression and classification problems can be solved using the K Nearest Neighbors (KNN) technique. This approach is suitable for all important parameters. It is widely used due to its fast interpretation and short computation time. No need to assume otherwise. It works well in multiclass situations. The ANN algorithm predicts similar things nearby. There are similar ones nearby. Compute distances using Euclidean distance functions. This method consists of

repeatedly running the KNN algorithm with different K values and choosing a K value that reduces the number of errors. Determine the ideal value of K by using cross-validation to measure precision or validation error.

A supervised learning technique known as the KNN can be applied to both classification and regression applications. In order to predict the label of an input data point, it first locates the k-nearest training set data points to the input data point. Then it uses the labels of those nearest neighbors.

Here are the steps to use the KNN algorithm:

- ✓ Choose the k nearest neighbor that will be utilized to produce the forecast.
- ✓ Measure the separation between the input data point and each training set data point.
- ✓ The k-nearest data points should be chosen depending on the distance determined in the previous phase.
- ✓ For classification tasks, select the majority label among the k-nearest neighbors to forecast the label of the input data point. Predict the value of the input data point for regression problems by averaging the values of the k-nearest neighbors.

### **3.2.7.2. Decision Tree Algorithm**

Developing prediction algorithms for target variables, or establishing classification systems based on numerous variables, are two prominent applications of decision tree techniques[12]. A decision tree is a type of tree structure where each leaf node represents a result and each inner node represents a property. Top-level nodes get the ability to split based on the value of an attribute. Recursive splitting is the process used to split a tree. Decision making is supported by this algorithm. A decision tree is a flowchart tree structure. Records in the database are divided into subsets by the values of one or more fields in the decision tree. This approach can be done recursively for each subset until all instances of each node are assigned to a single class.

Decision trees can handle category and numerical data and are simple to comprehend and interpret. However, if the tree is too complex, they could experience over fitting. Both classification and regression tasks can be performed using the decision tree technique. Until the

resulting subsets are as pure as is feasible, the data is separated recursively depending on the values of the characteristics.

### 3.3.8. Evaluation Metrics and Model Performance

Accuracy performance measure has used to predict and analyze rape crime. Accuracy is used to measure model performance. It is used to measure how the best performing model fits the data set. It measures how closely a value corresponds to reality (or the value agreed on and confirmed by many scientists). Accuracy is used to measure model performance. It is used to measure how the best performing model fits the data set. Precision ranges from 0 to 1. If the accuracy is close to one, the best performing model will fit the data set. Otherwise, the model will not fit the data set.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \text{ --- Eq(3.1)}$$

**Table 3.4: Confusion Matrix**

Actual	Predicted		
		Positive	Negative
	Positive	TP	FN
	Negative	FP	TN

- True Positive (TP): Overall percentage of members classified in class A belonging to class A.
- False Positive (FP): Total percentage of class A members who do not belong to class A.
- False Negative (FN): Overall percentage of class A members incorrectly classified as not belonging to class A.
- True Negative (TN): It does not classify the total percentage of members who do not belong to class A.

### 3.3.9. Implementation tools

This study uses multiple implementation tools and packages to implement a proposed rape crime prevention solution. Python programming language was used for implementing in each proposed solution through data preprocessing to the model building phase as well as for evaluating implementations and proposed classifier models. Python is used in this research because it is the programming language of choice for developers, researchers, and data scientists who need to work with data mining models. The following table provides a list of implementation tools and Python packages used in this study, along with their versions and descriptions.

**Table 3.5: Description of the tools and python packages used during the implementation**

No	Tool	Explanation
1	Python 3.9.0	A compile environment for python code easy to learn, powerful programming language to develop a machine learning application.
2	Sub line text 3	Easy cross-platform code editor well-known for its speed, comfort of use. It's an incredible editor right out of the box, but the real power comes from the ability to enhance its functionality using Package Control and creating custom settings.
3	Microsoft Excel 2013	Used data preparation tasks in cleaning, filtering, sorting the collected data, and remove duplicated data Also, used to manage the annotation task.
4	Scikit-learn	A set of python modules for machine learning and data mining. This study uses it for feature extraction and training and testing model. The name of the package is called sklearn.
5	Pandas	High-performance, easy-to-use data structures, and data analysis tools. This study uses it for data reading, manipulation, writing and handling the data frame.
6	NumPy	Array processing for number, strings, and objects.

## **CHAPTER FOUR**

### **RESULTS AND DISCUSSIONS**

#### **4.1. Introduction**

This chapter discussed about how preprocess the dataset and how the predictive model can be developed by using the preprocessed dataset. It also explained the evaluation metrics of the model using K-nearest neighbors (KNN) and Decision Tree (DT).

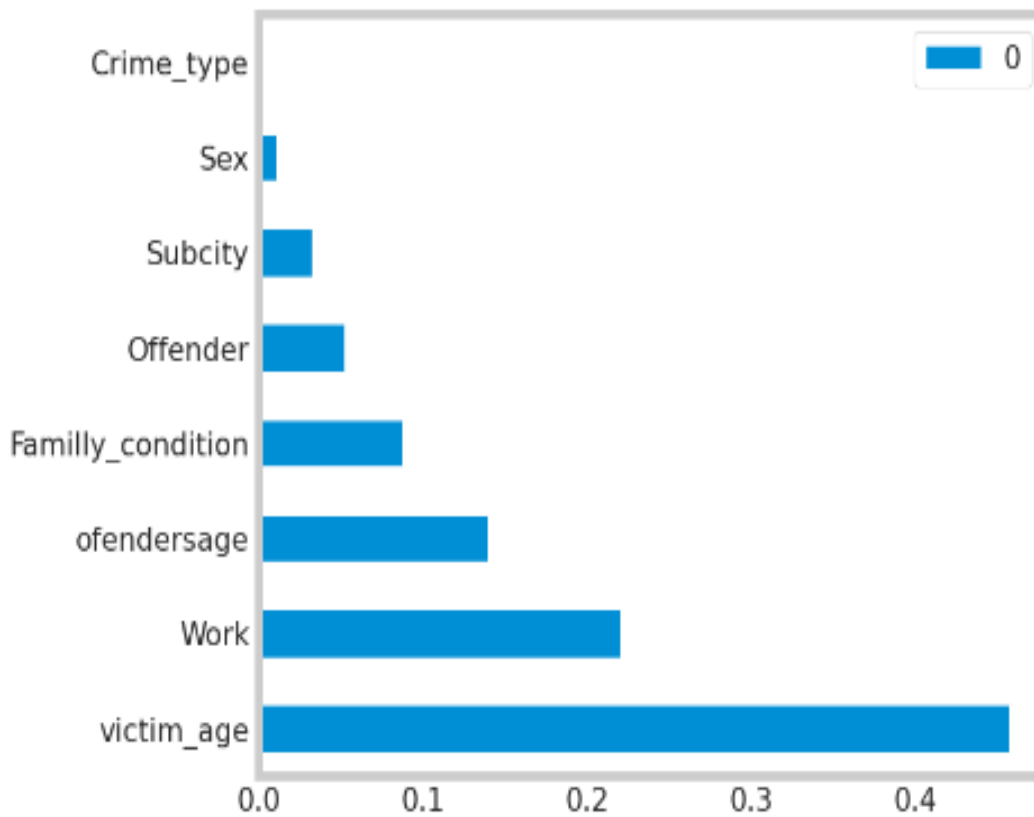
#### **4.2. Dataset**

The dataset which were used for doing this analysis were obtained from Addis Ababa police commission (AAPC). The AAPC collects data on crime victimization in all 11 sub-cities of Addis Ababa City. The dataset contains different attributes such as date, crime type, victim age, sex, grade level, work, family condition, offender, offenders age and sub city. It covered the year from 2006 to 2014 and included information on 4067 rows within 10 columns.

##### **4.2.1. Dataset Analysis and Feature Selection**

After preprocessing the data, the next step is to identify which features are most relevant for predicting rape crimes. Feature selection is the process of selecting a subset of relevant features from a larger set of features to use in a predictive model. This may involve using techniques such as correlation analysis, principal component analysis, or feature ranking methods to determine which features have the greatest impact on predicting the likelihood of a rape crime. The graph below shows the results of a feature selection analysis for the factors of date, victim age, sex, grade level, work, family condition, offender, offender's age, and sub city. The graph revealed the importance of each factor in predicting rape crime, with higher values indicating greater importance.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8ecf0b2130>
```



**Figure 4.1: Feature Selection**

The above feature selection graph revealed that victim's age has the highest value indicating that it is the most important factor that affects rape crime in the city. Offenders age and work type have also greater feature selection result. On the contrary, sex and sub city have the lowest value.

Once the most relevant features have been identified, the next step is to select an appropriate data mining model that can be used to predict rape crimes based on these features. This may involve using techniques such as decision trees, or KNN.

The `corr()` method calculates the relationship between each column in your data set.

```
#Using Pearson Correlation
plt.figure(figsize=(12,10))
cor = df.corr()

<Figure size 1200x1000 with 0 Axes>

[ ] #Correlation with output variable
cor_target = abs(cor["Subcity"])
#Selecting highly correlated features
relevant_features = cor_target[cor_target>0]
relevant_features

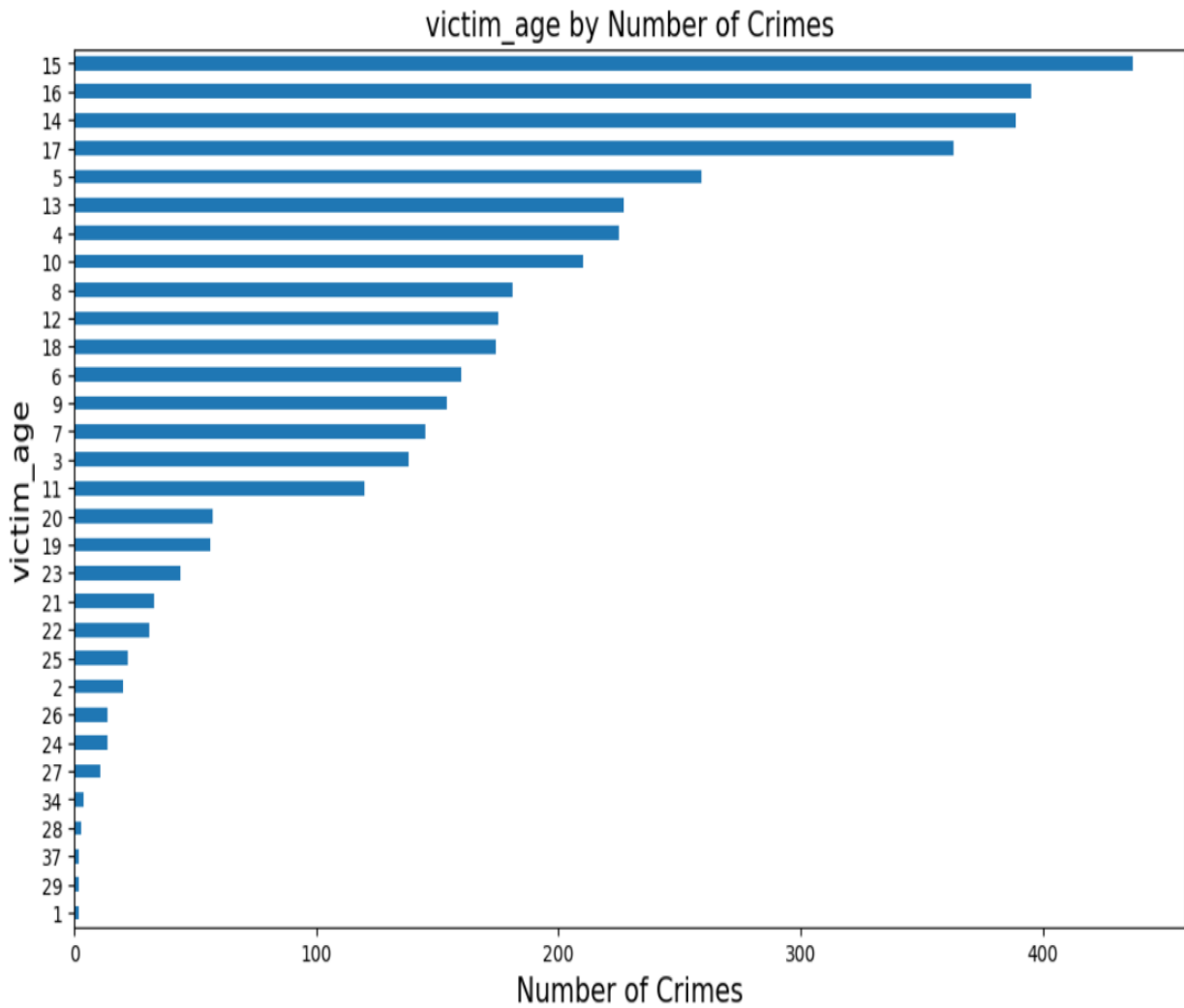
victim_age      0.001033
Sex              0.009682
Grade_level     0.001301
Work            0.008224
Family_condition 0.001596
Offender        0.008428
ofendersage     0.014674
Subcity         1.000000
month           0.014265
year            0.008428
Name: Subcity, dtype: float64
```

#### Figure 4.2: The correlation values of the dataset

The above graph revealed that, the most important factors for predicting rape crime are victim age, work type, offender's age, and family condition. These factors have the highest importance scores and should be included in the predictive model. Other factors have lower importance scores and are not relevant for predicting the model.

### 4.2.1.1. Victim Age and Rape Crime

The age of the victim is the most important factor in predicting and preventing rape. Younger victims may be more vulnerable to sexual assault, while older victims may be less likely to report the crime. The graph below shows the distribution of reported rape cases by age group.



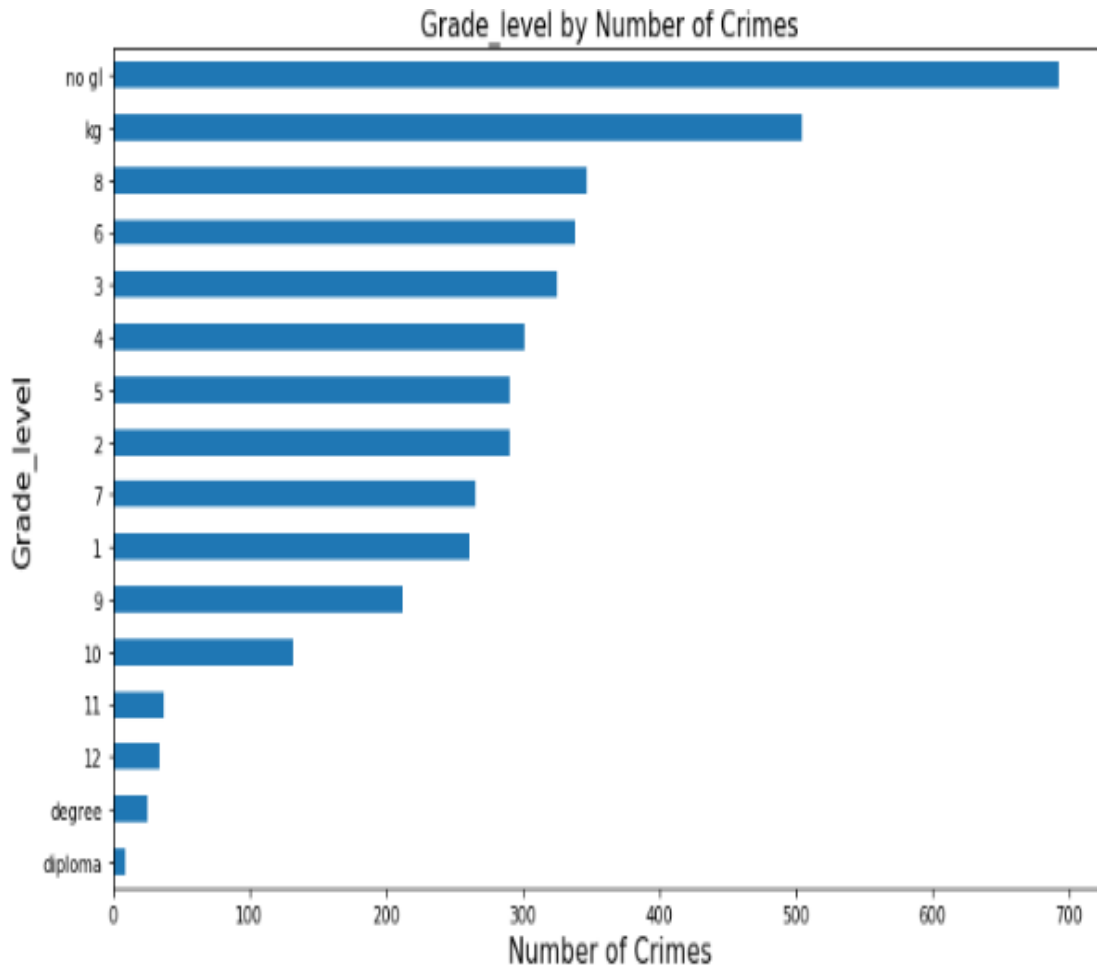
**Figure 4.3:Victim age by number of crime**

The above graph showed that the victim ages of 15, 16, 14, 17 and 5 years are the most affected age levels. On the contrary the age level of 1, 29, 37, and 34 years have less numbers of victims.



### 4.2.1.2. Education Status and Rape Crime

The education level of the victim can also be an important factor in predicting and preventing rape. The graph below shows the distribution of reported rape cases by grade level.



**Figure 4.4:Grade level by number of crime**

The above histogram revealed that no grades and kg students are highly victim in the rape crime. High school students are more vulnerable to sexual assault than elementary school students. On the contrary the analysis shows that the problem is not severe at kg level, diploma and degree holder.

### 4.2.1.3. Rape Crime and Sub cities

The problem of rape crime has highly varied across sub cities. The graph below indicated that Kolfe and Bole sub cities are sub cities where most of the rape crime victims are found. On the contrary Lemikura, Lideta and Akaki sub cities are the least affected areas.

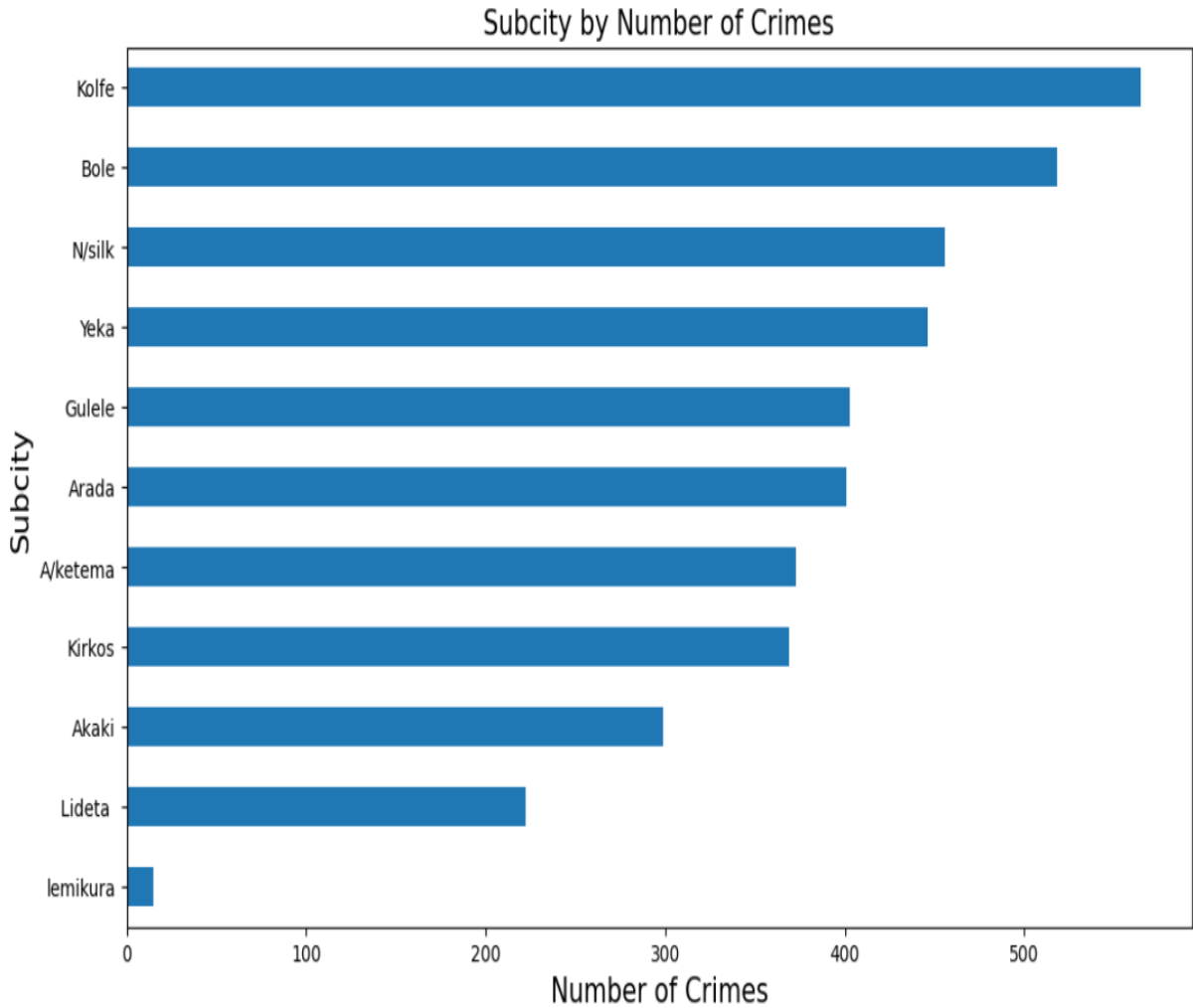


Figure 4.5: sub city by number of crime

#### 4.2.1.4. Rape Crime and Sex

The sex of the victim and offender is also an important factors in predicting and preventing rape. Women are more likely to be victims of rape than men. The graph below shows the distribution of reported rape cases by sex, with separate bars for male and female victims. The graph shows the majority of reported rape cases involve female victims. The ratio rape crime victim of the female more than the male. The female victim around 87.5 percent and the male is victim are around 12.5 percent. This information can be used to develop prevention strategies that are targeted at women, such as self-defense classes or increased awareness campaigns. There are also cases where male victims are involved though it is not severe like females. This information can also be used to develop prevention strategies that are targeted at men.

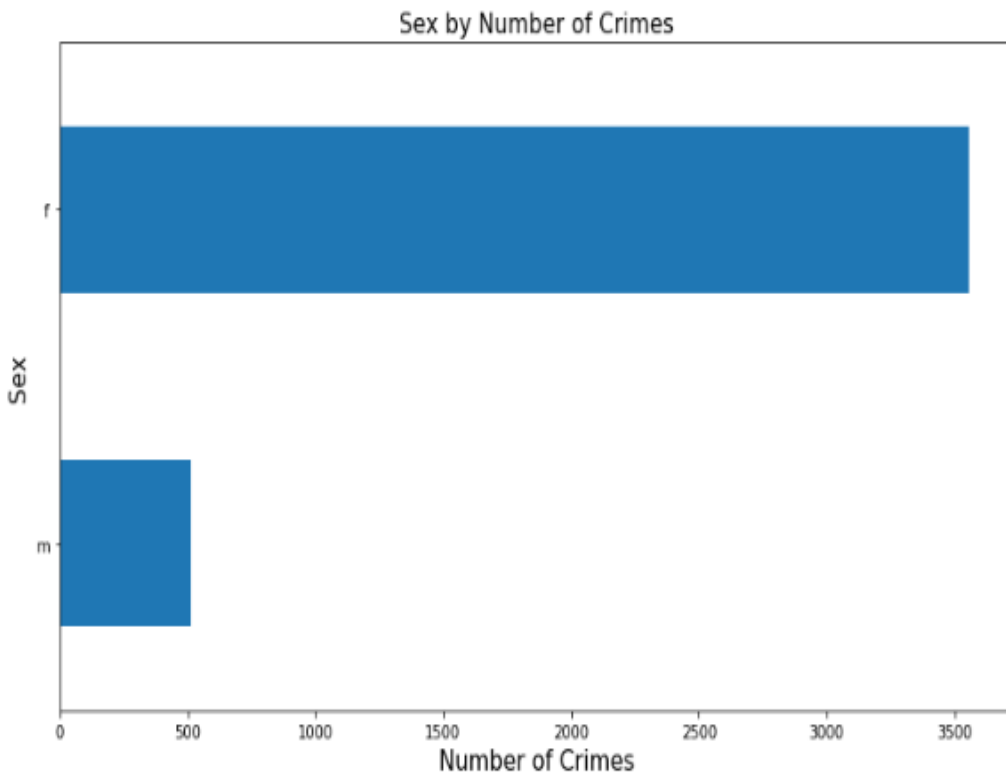


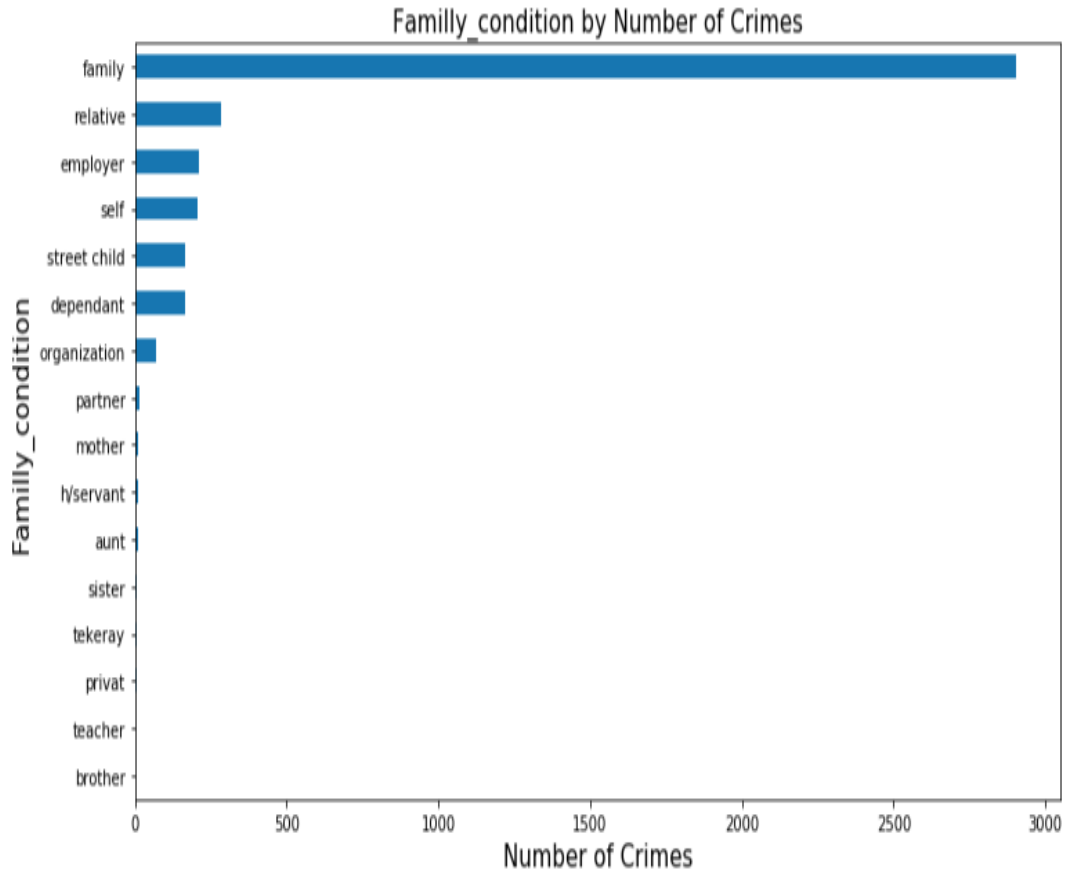
Figure 4.6:Sex by the number of crime

```
print('What are the percentage of fEMALE and MALE values in the dataset?')
pos=df[(df['Sex'] == "m")]
pos_percentage=len(pos)/len(df)
neg_percentage=1-pos_percentage
print('male instance percentage is ',pos_percentage)
print('Female instance percentage is ',neg_percentage)
```

```
What are the percentage of fEMALE and MALE values in the dataset?
male instance percentage is  0.12613720186869928
Female instance percentage is  0.8738627981313007
```

#### **4.2.1.5. Family Condition and Rape Crime**

The family condition of the victim can also be an important factor in predicting and preventing rape crime. The graph below shows the distribution of the victim's family condition in relation with the rape crime.

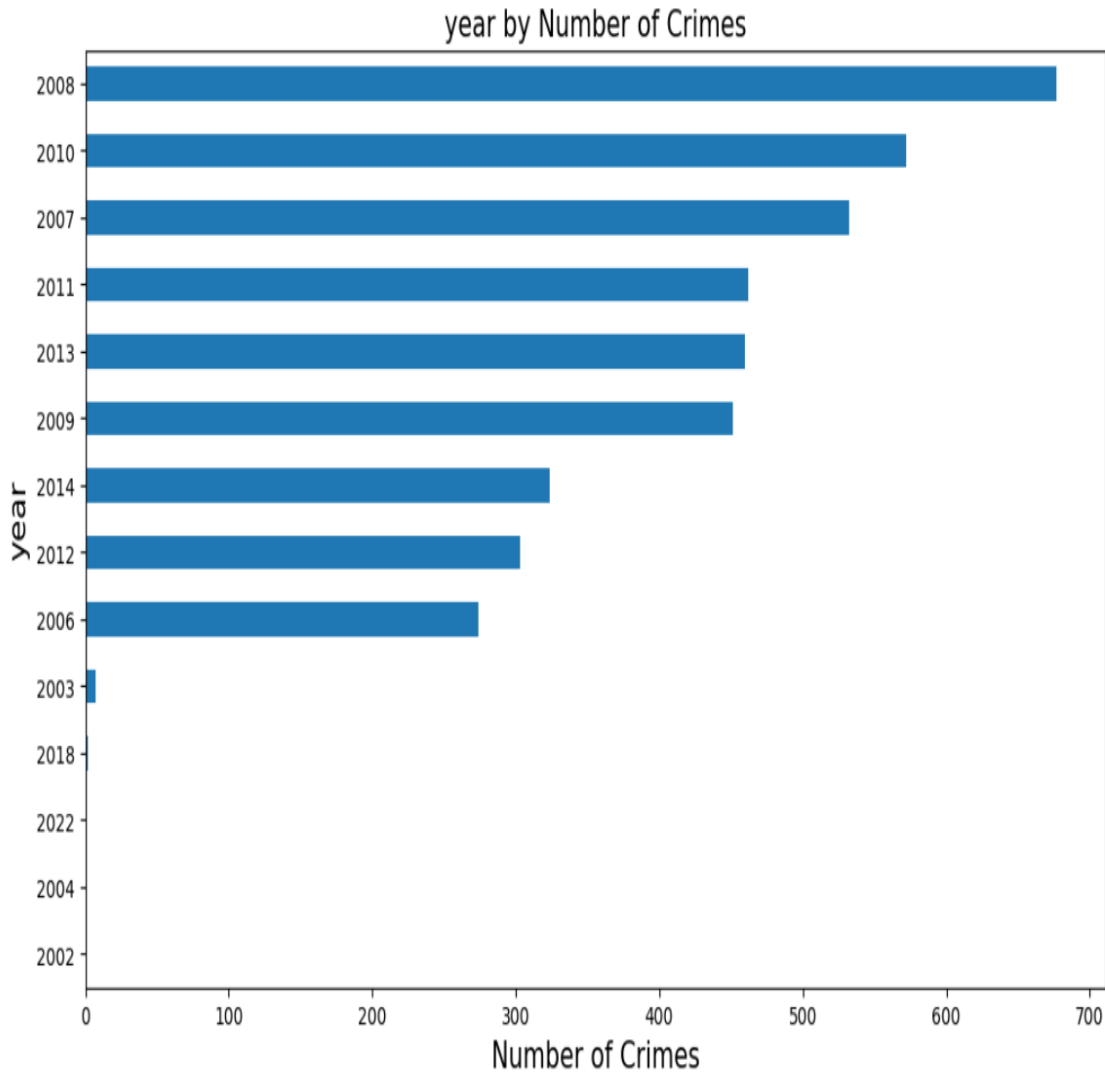


**Figure 4.7: Family condition and rape crime**

The above graph shows that most of the victims are living with their families. Moreover, many victims' are also living with relatives and employers. The others are living as street child and in Nongovernmental organizations (NGOs). By analyzing the data and identifying patterns and trends; we can develop a predictive model for rape crime prevention that takes into account the family ratio of the victim.

#### **4.2.1.6. Rape Crime and Years**

Severity of rape crime varied from year to year. The graph below shows the number of reported rape cases per year over a period of several years.



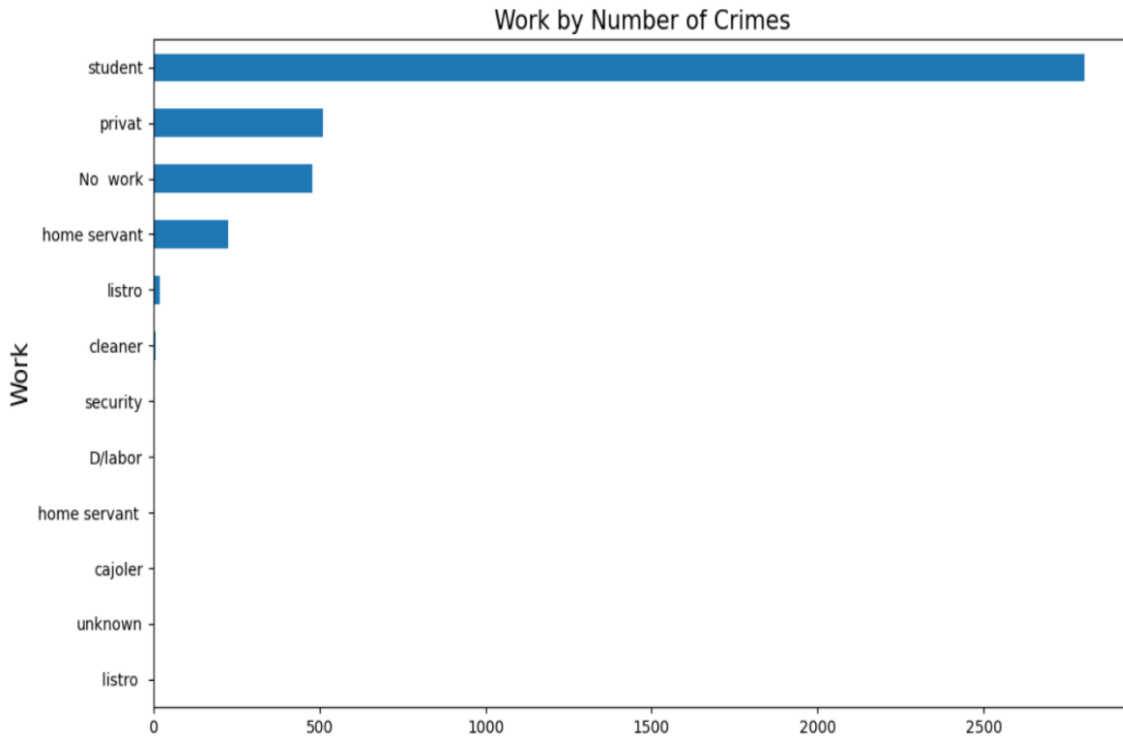
**Figure 4.8: Number of victim per a year**

The above graph revealed that the highest rape crime has occurred in crime occurred in 2008 followed by 2010,2007,2011,2013 and 2009.The year 2006, 2012and 2014 are years that a small amount of crime occurred. The graphical representation of the number of victims per year provides valuable insights into the occurrence of rape crimes and can be used to develop effective strategies for preventing sexual assault.

#### **4.2.1.7. Occupation and Rape Crime**

The occupation of the victim or offender can also be an important factor in predicting and preventing rape. The graph below shows the distribution of reported rape cases by

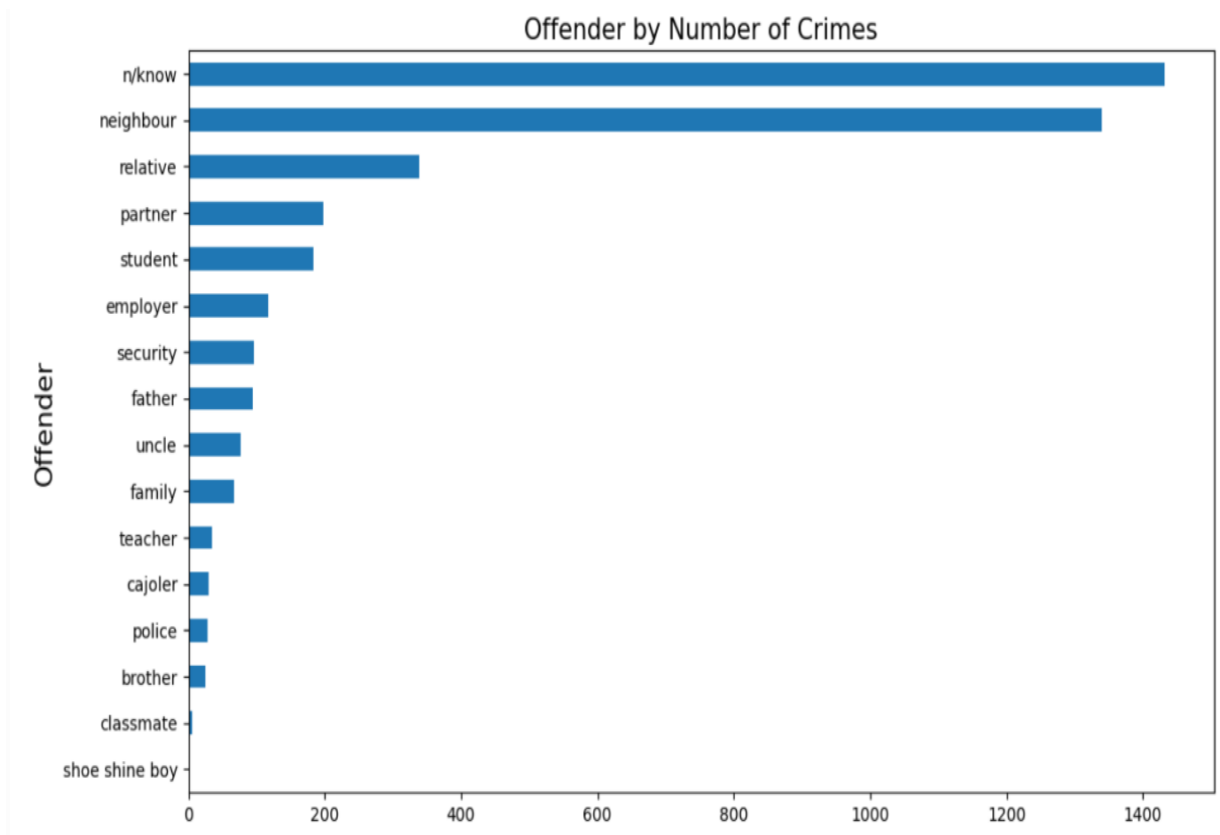
occupation. The graph shows students are highly assaulted in rape crimes. Moreover, a person who works as a home servant and a person who works in private institutions are also highly affected by rape crime. Those who are unemployed are also victims.



**Figure 4.9: Rape crime and occupation**

#### 4.2.1.8. Offenders and Rape Crime

The identity of the offender can be an important factor in predicting and preventing rape. Certain individuals may be more likely to commit rape based on their past criminal history or behavioral patterns. The graph below shows that most of the offenders are not known. Many of the victims are offended by their Neighbors and relatives.

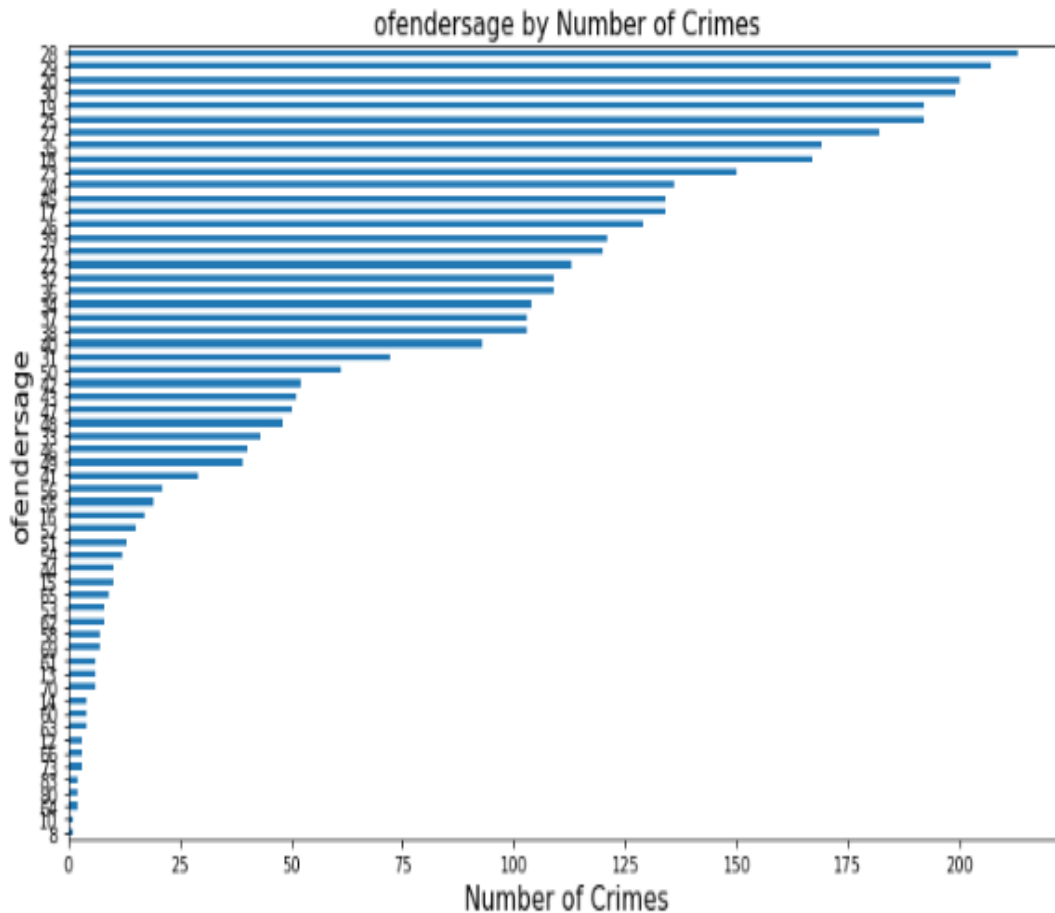


**Figure 4.10: Offenders and Rape Crime**

#### 4.2.1.9. Offenders Age and Rape Crime

Offenders' age is also the other most important factor in rape crime. Age distribution of offenders is one way to explore the relationship between offender age and rape crime prevention and prediction.





**Figure 4.11: Offenders Age and Rape Crime**

The above graph revealed that most of the offender ages are 28 and 29 years old. The ages of 20, 30 and 19 years are also among the highest age of offenders.

### 4.3. Model Development

K-Nearest Neighbors (KNN) and Decision Tree (DT) data mining techniques have been used for developing predictive modeling. Both techniques can be used to develop a predictive model for rape crime prevention. The first step in developing a KNN model is to select the value of K, which is the number of nearest neighbors to consider when making a prediction. The selected features are then used to calculate the distance between each data point in the training set and the new data point. The K nearest neighbors then identified, and the outcome of the new data point is

predicted based on the outcomes of the K nearest neighbors. Secondly, to develop the DT model, select the root node, which is the feature that best splits the data into two groups. The data is then split based on the selected feature, and the process is repeated for each subset of data until a stopping criterion is met. The outcome of the new data point is predicted based on the path through the decision tree.

Generally, the KNN and DT model development process for developing a predictive model for rape crime prevention using data mining techniques involves selecting the appropriate value of K or the root node, calculating the distance or splitting the data based on the selected feature, predicting the outcome based on the K nearest neighbors or decision tree path, evaluating the performance of the models, and using the models to identify potential risk factors and develop prevention strategies.

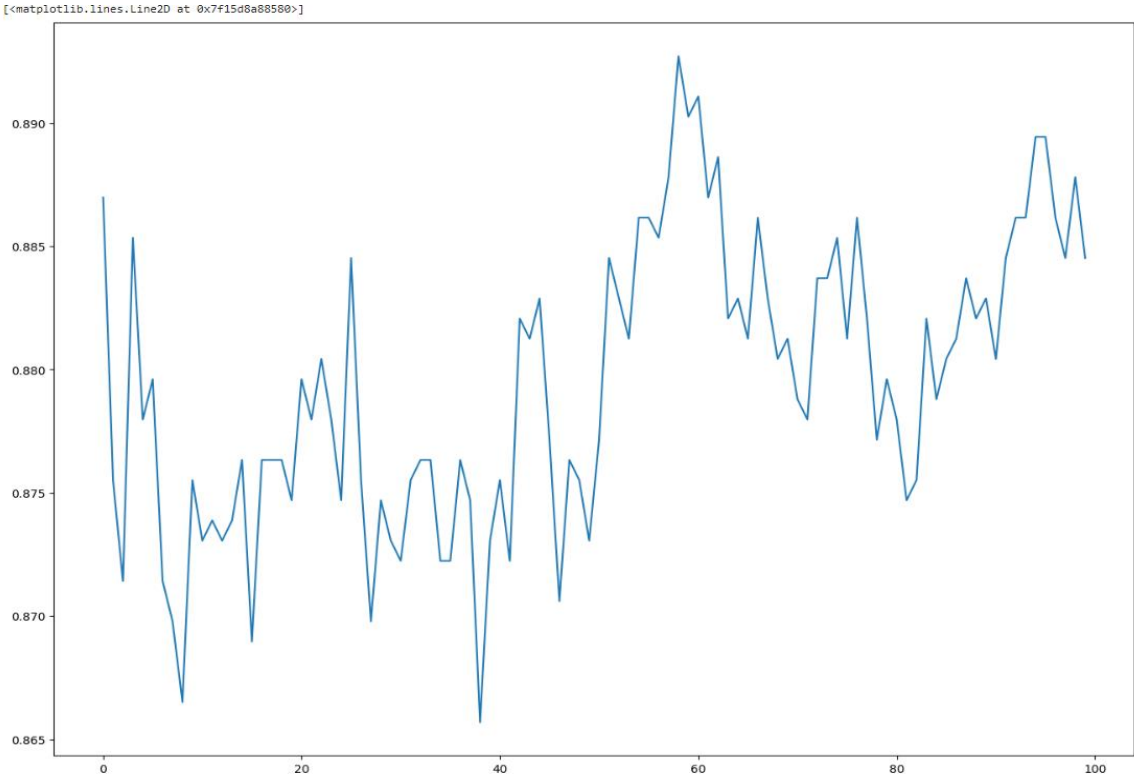
The dataset present in CSV format has pre-processed and ready to train the model. The data set can be split into 80% for training and 20% for testing. After been developed and tested, the model can be used to forecast and prevent rape offences. This can mean integrating the model into already-existing crime prevention initiatives or using the model to identify high-risk populations or locations that require more support or intervention. By examining several factors like demographics, location, date, and other parameters, KNN and DT models can be used to predict the probability of rape crimes in a given area. These models can be used to clarify the reasons of rape incidents and offer recommendations for their avoidance.

To process the data using these algorithms, first the algorithm which is more suitable for the problem should be determined. Then, preprocess the data by cleaning, transforming, and splitting it into training and testing sets. Finally, train the chosen algorithm on the training set and evaluate its performance on the testing set. There are many libraries available in different programming languages, such as Scikit-learn in Python that can support to implement these algorithms.

#### **4.3.1. K-Nearest Neighbors Algorithm Implementation**

KNN is a particular kind of algorithm used to solve classification issues. In KNN, the majority class of a new instance's k nearest neighbors is used to predict it. The dataset and the current challenge both influence the choice of K. KNN is a straightforward and efficient prediction

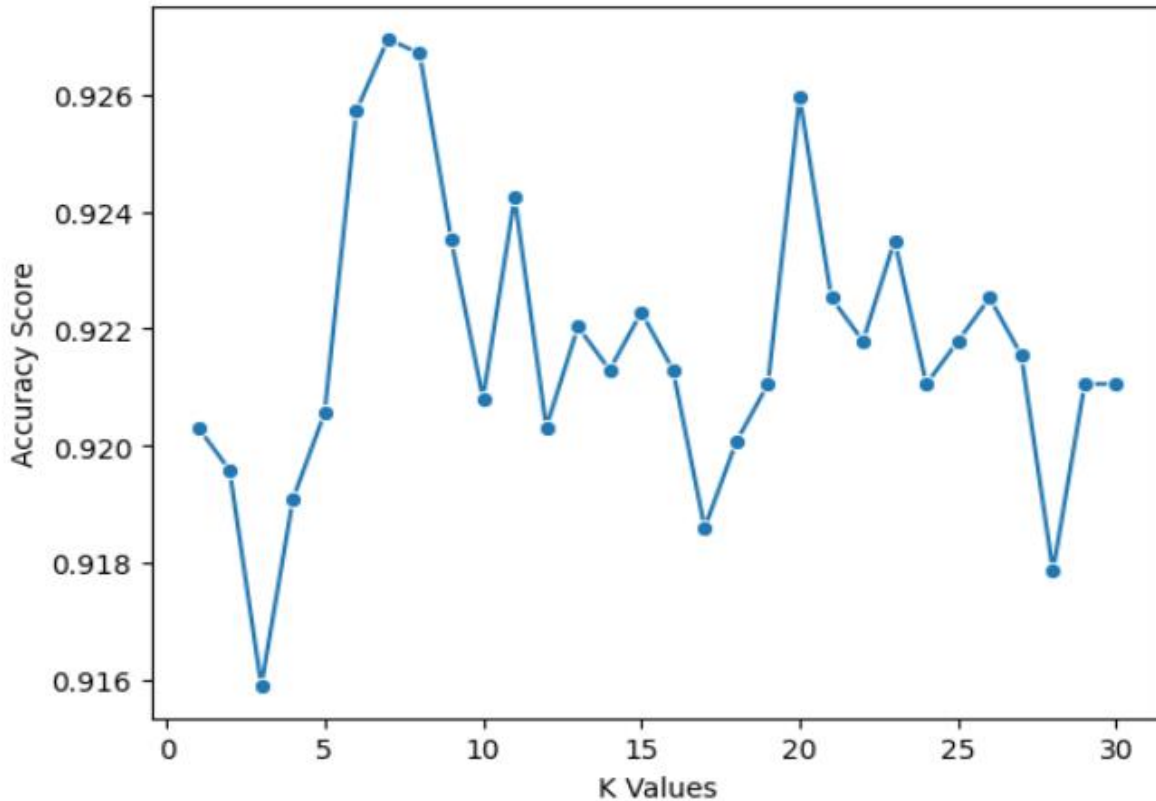
approach, however, it may struggle with high-dimensional or poorly segregated data. KNN algorithms use feature similarity to predict the value of every new data point.



**Figure 4.12:Error rate**

The error rate in the KNN model refers to the proportion of misclassified instances in the testing set. In other words, it is the percentage of cases where the model predicted the wrong outcome. It is important to minimize the error rate in the KNN model to ensure that the model is accurately predicting the likelihood of rape crime based on the selected features. This can be achieved by selecting the appropriate value of K, selecting relevant features, and ensuring that the training and testing sets are representative of the dataset. The error rate in the KNN model is also an important metric for evaluating the performance of the model and ensuring that it is accurately predicting the outcomes of reported rape cases.

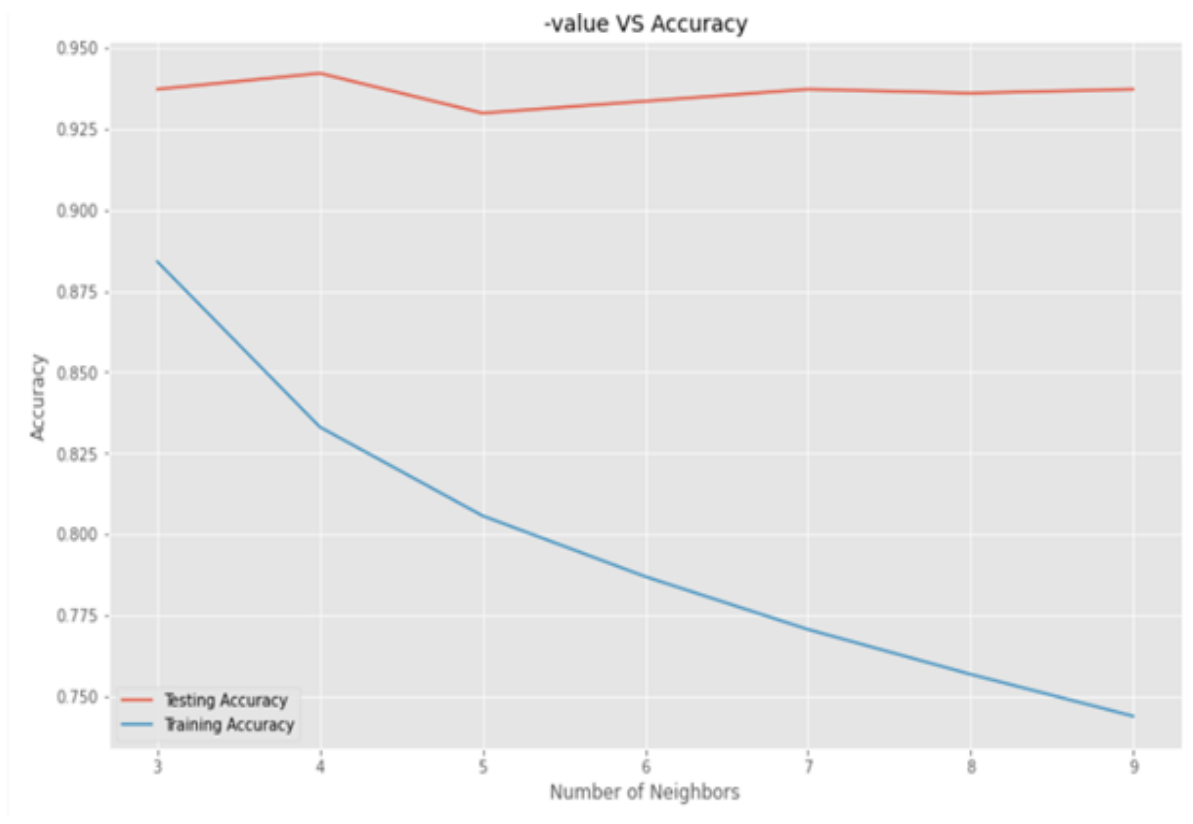
The graph below shows the accuracy of the KNN model for different values of K. This data can be used to identify the optimal value of K for the KNN model.



**Figure 4.13: Selection of K values**

The graph shows that, the accuracy of the KNN model varies depending on the value of K. The highest accuracy is achieved when K is equal to 7, with an accuracy of approximately 93 percent. This information can be used to select the optimal value of K for the KNN model. By selecting the appropriate value of K, we can improve the accuracy of the model and ensure that it accurately predicts the likelihood of sexual assault based on the selected features. The graphical representation of the selection of K values in the KNN model provides valuable insights into the performance of the model and can be used to develop an effective predictive model for rape crime prevention using data mining techniques.

The graph below shows the train-test accuracy of the KNN model for different values of K. This data can be used to evaluate the performance of the KNN model and identify the optimal value of K.

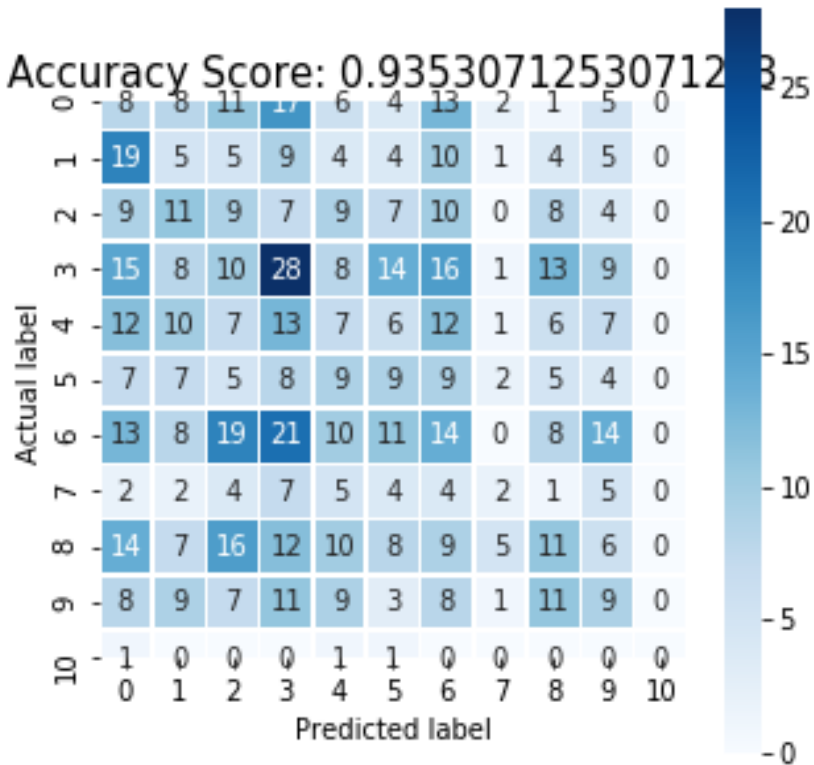


**Figure 4.14: Train Test Accuracy in KNN**

The above graph indicated that the train-test accuracy of the KNN model varies depending on the value of K. The highest train-test accuracy is achieved when K is equal to 4, with an accuracy of approximately 0.93 in train test accuracy graph. This information can be used to select the optimal value of K for the KNN model. By selecting the appropriate value of K, we can improve the accuracy of the model and ensure that it accurately predicts the likelihood of sexual assault based on the selected features. This train-test accuracy result in the KNN model provides valuable insights into the performance of the model and can be used to develop an effective predictive model for rape crime prevention using data mining techniques.

The graph below shows a graphical representation of the confusion matrix for the KNN model.

Text(0.5, 1, 'Accuracy Score: 0.9353071253071253')



**Figure 4.15:Confusion Matrix of KNN Model**

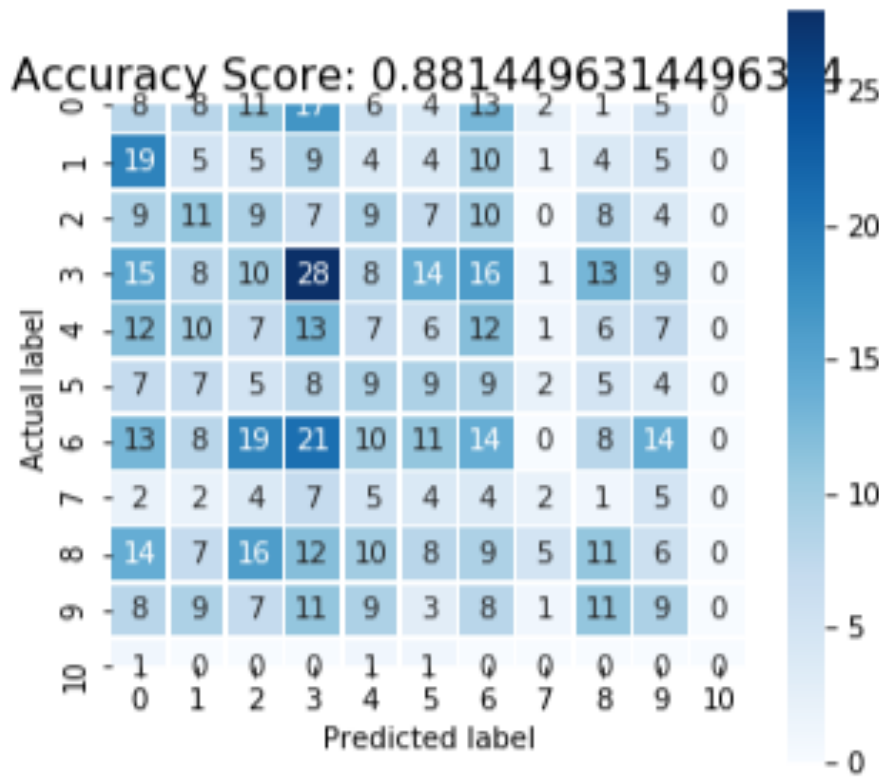
The above graph shows that the KNN model correctly predicted 93.5% of the reported rape cases. The confusion matrix provides valuable insights into the performance of the KNN model and can be used to evaluate the accuracy of the model. By analyzing the confusion matrix, we can identify potential areas for improvement in the model and develop strategies for improving its accuracy. Overall, the graphical representation of the confusion matrix for the KNN model provides valuable insights into the performance of the model and can be used to develop an effective predictive model for rape crime prevention using data mining techniques.

### 4.3.2. Decision Tree Implementation

Decision trees are applied to classification issues. In decision trees, the values of the characteristics are used to divide the data into smaller subsets. Next, based on the values of yet another attribute, each subset is further divided, and so forth.

The graph below shows the confusion matrix for the Decision Tree model.

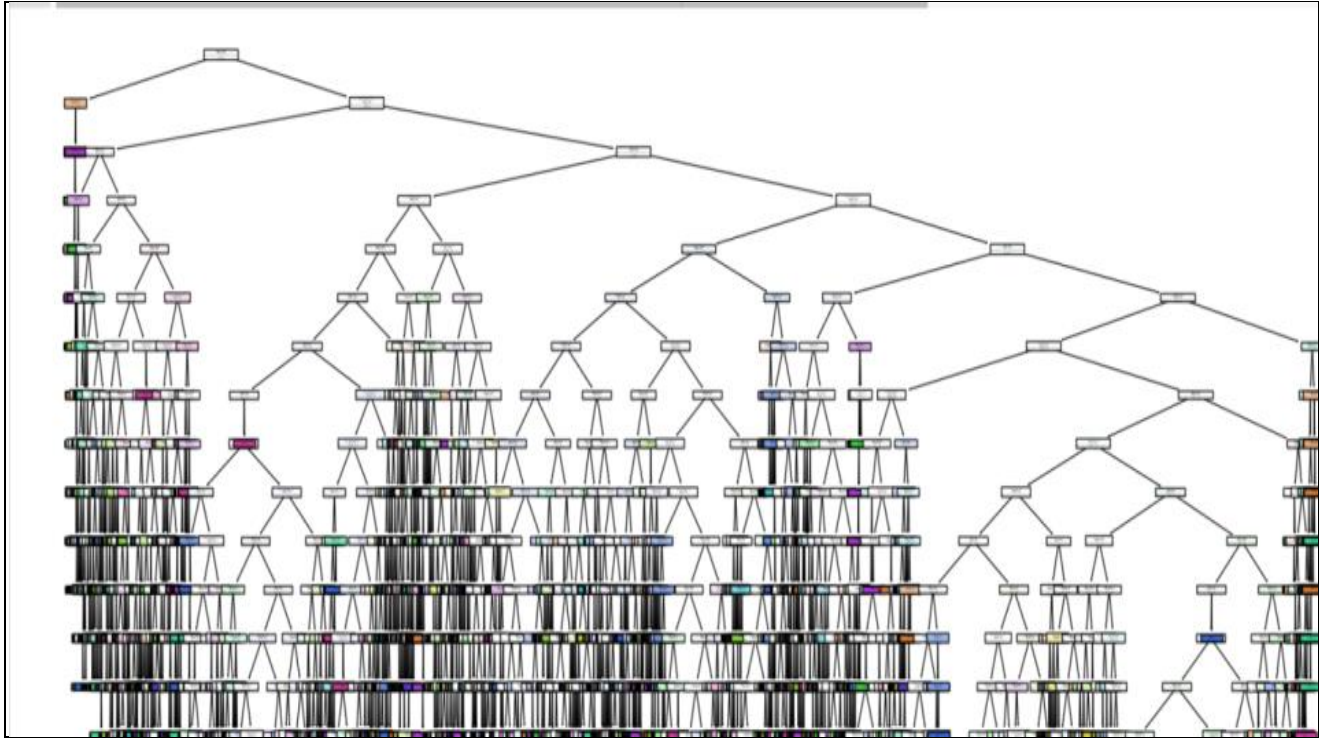
`Text(0.5, 1, 'Accuracy Score: 0.8814496314496314')`



**Figure 4.16: Visualizing the Decision Tree Confusion Matrix**

The above graph revealed that the decision tree model correctly predicted 88.1% of the reported rape cases. The confusion matrix provides valuable insights into the performance of the decision tree model and can be used to evaluate the accuracy of the model. By analyzing the confusion matrix, we can identify potential areas for improvement in the model and develop strategies for improving its accuracy. The graphical representation of the confusion matrix for the Decision

Tree model provides valuable insights into the performance of the model and can be used to develop an effective predictive model for rape crime prevention using data mining techniques.



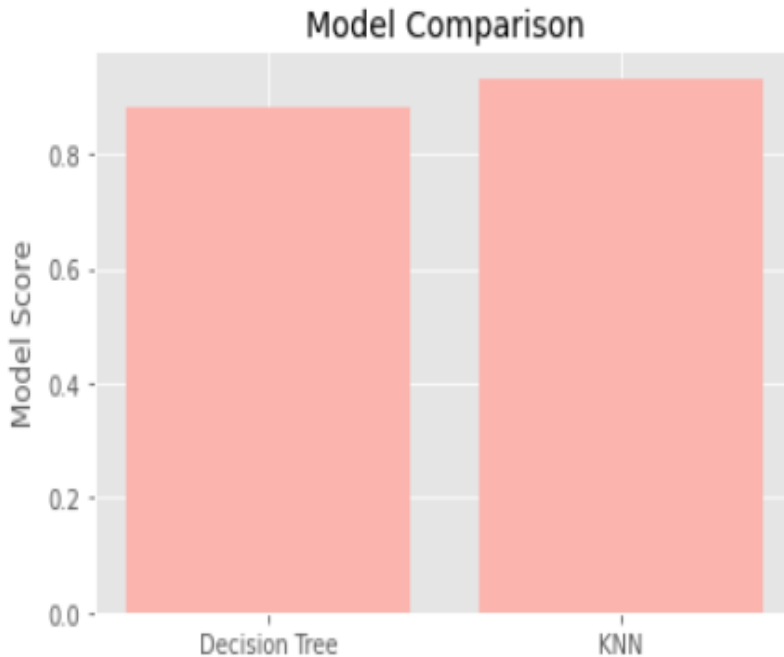
**Figure 4.17: Decision tree structure**

#### **4.4. ModelEvaluation**

The model evaluation process involves assessing the performance of the KNN and DT models using various metrics, such as accuracy. These metrics provide a measure of how well the models are performing in terms of correctly predicting the outcomes of reported rape cases. To evaluate the performance of the KNN and DT models, first split the data into training and testing sets. The training set is used to train the models, while the testing set is used to evaluate their performance. Once the models have been trained, their performance will be evaluated using the testing set. The calculated the accuracy result of the model is used to compare them to identify the best-performing model.



Different parameters can be analyzed and used to calculate the performance of the model that fits the dataset. The results obtained from the proposed models are visualized in the following figure:



**Figure 4.18: Performance of Prediction Model**

The above figure shows the performance of the model in model comparison. The model score indicated that KNN provides 93.5% accuracy, and decision trees provide 88.1% accuracy. Therefore, KNN provides better prediction as well as analysis of rape crime.

#### **4.5. Result and Discussions**

The result discussion involves analyzing the performance of the KNN and DT models and also discussing the implications of the results for rape crime prevention. The result discussion is an important step in developing an effective predictive model for rape crime prevention using data mining techniques. By analyzing the performance of the models and discussing the implications of the results, it is possible to develop strategies for improving public safety and preventing rape crime.

The training error can be obtained for the same data that the model is trained with, and the accuracy can be obtained for the data set aside for testing. The training error indicates how well the model fits the data, and the test error shows how well the network can generalize to unseen data samples. The model's performances have been measured using error rates and accuracy. The performance of crime prediction under the experimental analysis in data mining techniques, besides the model in crime prediction, has been developed using different models like KNN and decision trees, and we have gotten performance of 93.5% and 88.1% of accuracy respectively.

From the analysis, the k nearest neighbor fits the data with better performance. The performance of prediction is better in the KNN than in other models particularly DT method. For the comparison of models of rape crime prediction, we have to use accuracy as a performance measure, and we have a higher performance and a larger dataset compared with the previous paper. The existing research predicts and analysis the crime data mining techniques with a small amount of data and a few attributes. This thesis used a larger dataset with various parameters when using a data mining techniques. The number of models in the crime prediction and analysis is not limited. The existing research considers different parameters. The existing research has the lowest performance model with the fewest attributes. The state of art in the prediction and analysis of crime has developed by using KNN and decision tree. Most of the existing paper does not determine the parameters and the researchers do not normalize the dataset. The data set has normalized, evaluated the model with a validation set, and optimized by using K-fold cross-validation.

On the developing a predictive model for rape crime prevention using data mining techniques such as KNN and DT to be very interesting and informative, the study provided valuable insights into the use of data mining techniques for crime prevention. The researchers used appropriate data mining techniques, including KNN and DT algorithms, and evaluated the performance of the models using appropriate metrics. The study also included a thorough description of the data sources and preprocessing techniques used, which helped to ensure the accuracy and reliability of the results. The key findings of the study were that the KNN and DT models were effective in predicting rape crime and identifying high-risk sub city in Addis Ababa. One limitation of the study was that it relied on historical data, which may not be representative of current or future trends in rape crime. Additionally, the study did not address the ethical considerations involved

in using predictive models for sensitive and personal crimes like rape, which is an important area for future research. The study has important implications for law enforcement agencies and policymakers, as it highlights the potential of predictive models for identifying high-risk periods and allocating resources. However, it is important to approach this type of analysis with sensitivity and care, and to ensure that any predictive models developed are used ethically and do not achieve biases or discrimination. Policymakers should work to build trust and transparency with the communities they serve, and ensure that any data used in the analysis is obtained and used in a responsible and ethical manner.

#### **4.6. User Acceptance Evaluation**

User acceptance evaluation has been conducted so as to check model validity. The paper has reviewed by various stakeholders such as police commission employees, individuals, and IT experts and they give some ideas about the model development of the rape crime prediction. They suggested that the model needs to be more clearly develop or better organized to make the research findings more accessible to users. They provide feedback on the research methodology, suggesting improvements or pointing out potential flaws in the study design. They also provide feedback on the statistical analysis of the data, suggesting alternative methods or pointing out potential issues with the analysis. They commented on the significance of the model, suggesting ways to better contextualize the results or highlighting potential implications for future research.

## CHAPTER FIVE

### CONCLUSION AND RECOMMENDATION

#### 5.1. Conclusion

The issue of rape crime is a very sensitive and serious issue, and data mining techniques can be used to analyze and predict rape crimes to prevent them from happening in the future. However, it is important to note that data mining techniques should not be used to blame victims or to excuse the actions of perpetrators. Instead, they should be used to help identify patterns and risk factors associated with rape crimes, so that put preventative measures in place to reduce the likelihood of such crimes occurring. The detected determinants of rape crime include victim's age, offender's age, occupation, gender, location, and relationship to the offender. KNN and DT data mining techniques have been used so as to develop prediction model of rape prevention. Developing a predictive model for rape crime prevention using KNN and DT data mining techniques involves selecting the appropriate value of K or the root node, calculating the distance or splitting the data based on the selected feature, predicting the outcome based on the K nearest neighbors or decision tree path. The KNN and DT models were able to predict the probability of a rape crime occurring with an accuracy of 93.5% and 88.1% respectively. After the models have been trained, their performances have been evaluated using the testing data set and accuracy. The results show that the system is able to accurately predict the rape crime in a given area. The models can be used to identify areas where prevention efforts should be focused and also develop interventions to prevent rape crime.

#### 5.2. Recommendation

Based on the result of the study, the following recommendations are forwarded:

- The result revealed that degrees of rape crime vary among sub cities. Kolfe Keranya, Bole and Nifase Silik sub cities are the most affected areas. Therefore, the concerned body should give due attention in allocating resources and take precautionary action so as to prevent rape crime in the city.
- Those individuals whose ages are between 14 and 17 years are highly affected so that the law enforcement agencies and others should give attention for these age groups.

- The result of the paper indicated that those individuals who are illiterate and students at KG level are highly exposed for rape crime. Therefore, special intervention mechanism should be designed and applied in these areas.
- The governmental and non-governmental organizations should use the result of this paper as an input for developing rape crime prevention policies and strategies.
- The analysis of rape crime should be conducted based on different parameters with different dataset types. Therefore, it is better for future studies to use large data set for the prediction of rape crime occurrence. Moreover, taking a combined approach of time series prediction and other data mining techniques to make a hybrid model for higher performance.
- Future studies should investigate the moral issues raised by using predictive models for sensitive and private crimes like rape and propose policies for their appropriate application.

## References

- [1] G. Saltos and M. Cocea, "An exploration of crime prediction using data mining on open data," *Int. J. Inf. Technol. Decis. Mak.*, vol. 16, no. 05, pp. 1155–1181, 2017.
- [2] S. Spoo et al., "Victims' attitudes toward sex offenders and sex offender legislation," *Int. J. Offender Ther. Comp. Criminol.*, vol. 62, no. 11, pp. 3385–3407, 2018.
- [3] P. Karazivan et al., "The patient-as-partner approach in health care: a conceptual framework for a necessary transition," *Acad. Med.*, vol. 90, no. 4, pp. 437–441, 2015.
- [4] R. Baker, "Data mining for education," *Int. Encycl. Educ.*, vol. 7, no. 3, pp. 112–118, 2010.
- [5] H. Letezgi, "Mining crime data for effective resource allocation and crime prevention: The case of Addis Ababa police commission," *Addis Ababa Univ.*, 2011.
- [6] J. Han and M. Kamber, "Classification and prediction," *Data Min. Concepts Tech.*, vol. 2006, pp. 347–50, 2006.
- [7] P. Prasad, R. Singh, and R. Kothari, "Crime analysis and prediction using data mining," *Int. Res. J. Eng. Technol.*, 2020.
- [8] I. H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques with Java implementations," *Acm Sigmod Rec.*, vol. 31, no. 1, pp. 76–77, 2002.
- [9] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.
- [10] S. Sathyadevan, M. S. Devan, and S. S. Gangadharan, "Crime analysis and prediction using data mining," in *2014 First international conference on networks & soft computing (ICNSC2014)*, IEEE, 2014, pp. 406–412.
- [11] S. Zahran, E. M. Mohamed, and H. M. Mousa, "Detecting and Predicting Crimes using Data Mining Techniques: Comparative Study," *IJCI Int. J. Comput. Inf.*, vol. 8, no. 2, pp. 57–62, 2021.
- [12] Y.-Y. Song and L. U. Ying, "Decision tree methods: applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, p. 130, 2015.
- [13] R. K. Sharma and S. Kumar, "Performance modeling in critical engineering systems using RAM analysis," *Reliab. Eng. Syst. Saf.*, vol. 93, no. 6, pp. 913–919, 2008.
- [14] K. Un, "Patronage politics and hybrid democracy: Political change in Cambodia, 1993-2003," *Asian Perspect.*, vol. 29, no. 2, pp. 203–230, 2005.

- [15] R. Rajamanickam, M. Z. M. Zahir, N. K. Dahlan, H. B. M. Saad, H. Hashim, and D. Nagarajan, "Analysis Of Criminal Offences Involving Teachers In Malaysia: 10.2478/bjlp-2022-007096," *Balt. J. Law Polit.*, vol. 15, no. 7, pp. 1323–1336, 2022.
- [16] M. NOGAJ, "Combatting Violence against Women: European Added Value Assessment (+ Annexes I-II)," 2013.
- [17] J. DeLamater, J. S. Hyde, and M.-C. Fong, "Sexual satisfaction in the seventh decade of life," *J. Sex Marital Ther.*, vol. 34, no. 5, pp. 439–454, 2008.
- [18] H. E. Jackson and M. J. Roe, "Public and private enforcement of securities laws: Resource-based evidence," *J. Financ. Econ.*, vol. 93, no. 2, pp. 207–238, 2009.
- [19] S. Hossain, A. Abtahee, I. Kashem, M. M. Hoque, and I. H. Sarker, "Crime prediction using spatio-temporal data," in *Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1*, Springer, 2020, pp. 277–289.
- [20] E. Runarsdottir, E. Smith, and A. Arnarsson, "The effects of gender and family wealth on sexual abuse of adolescents," *Int. J. Environ. Res. Public. Health*, vol. 16, no. 10, p. 1788, 2019.
- [21] S. Sorial and J. Poltera, "Rape, women's autonomy and male complicity," *Women Violence Agency Vict. Perpetrators*, pp. 15–33, 2015.
- [22] Khan et al., "Development of a Predictive Model for Rape Crime Prevention Using Data Mining Techniques," p. 2014.
- [23] M. Khan, A. Ali, and Y. Alharbi, "Predicting and preventing crime: A crime prediction model using san francisco crime data by classification techniques," *Complexity*, vol. 2022, 2022.
- [24] P. Yerpude, "Predictive modelling of crime data set using data mining," *Int. J. Data Min. Knowl. Manag. Process IJDKP Vol*, vol. 7, 2020.
- [25] S. Kim, P. Joshi, P. S. Kalsi, and P. Taheri, "Crime analysis through machine learning," in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, IEEE, 2018, pp. 415–420.
- [26] E. T. Castro and A. A. Hernandez, "Developing a predictive model on assessing children in conflict with the law and children at risk: a case in the Philippines," in *2019 IEEE 15th*

International Colloquium on Signal Processing & Its Applications (CSPA), IEEE, 2019, pp. 243–248.

- [27] D. K. Tayal, A. Jain, S. Arora, S. Agarwal, T. Gupta, and N. Tyagi, “Crime detection and criminal identification in India using data mining techniques,” *AI Soc.*, vol. 30, pp. 117–127, 2015.



## **Appendix**

The following questions have been distributed for stakeholders so as to conduct user's acceptance valuation.

1. What did you think about the model on developing a predictive model for rape crime prevention using data mining techniques such as KNN and DT?
2. What were the key findings of the model?
3. What were the limitations of the model?
4. What are the implications of the model for law enforcement agencies and policymakers?