HAWASSA UNIVERSITY

INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

PREDICTION AND ANALYSIS OF CRIME AGAINST WOMEN USING MACHINE-LEARNING

BY

VERONICA TESFAYE BELAYNEH

HAWASSA UNIVERSITY, HAWASSA, ETHIOPIA

MAY, 2024

PREDICTION AND ANALYSIS OF CRIME AGAINST WOMEN USING MACHINE-LEARNING

BY

VERONICA TESFAYE

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF HAWASSA UNIVERSITY INSTITUTE OF TECHNOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

HAWASSA, ETHIOPIA

Program: MSc. Computer Science

Major Advisor: Varaganithan Anitha (PhD)

Co-Advisor: Teshager Kassa (MSc)

HAWASSA UNIVERSITY

INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES FACULTY OF INFORMATICS

THESIS APPROVAL SHEET-I

This is to certify that the thesis entitled "Prediction and Analysis of Crime against Women Using Machine learning," submitted in partial fulfillment of the requirements for the degree of Master of Science with a specialization in Computer Science in the Graduate Program of the School of Informatics, has been conducted by Veronica Tesfaye Belayneh. We confirm that the student has met all the requirements and recommend that this thesis be submitted to the department.

Name of Major Advisor: Varaganithan Anitha (PhD)

Signature: _____

Date of submission: _____

Name of Co-advisor: Teshager Kassa (MSc)

Signature: _____

Date of Submission: _____

THESIS APPROVAL SHEET-II

This document is a certificate that confirms the thesis written by Veronica Tesfaye Belayneh, titled "Prediction and Analysis of Crime against Women Using Machine Learning- the Case of Hawassa City," which meets the standards set by the University in terms of originality and quality. This thesis adheres to the University's standards regarding originality and excellence. It fulfills the criteria for the Master of Science in Computer Science degree. The program's requirements align with the University's regulations and uphold recognized benchmarks for originality and quality.

Name of Major Advisor	Signature	Date
Varaganithan Anitha(PhD)	_____	_____
Name of Internal Examiner -I	Signature	Date
Degif Teka (PhD)	_____	_____
Name of Internal Examiner –II	Signature	Date
Mulat Shiferaw(MSc)	_____	_____
Name of External Examiner	Signature	Date
Mesay Adinew (PHD)		7/2/2024
SGS Approval	Signature	Date
	_____	_____

ACKNOWLEDGMENT

I am sincerely thankful to God for granting me the strength, wisdom, and time to accomplish this thesis. My appreciation extends to all my instructors during the MSc program for providing the essential courses that were fundamental in shaping the mindset and knowledge necessary for this work.

I express my heartfelt gratitude to my esteemed advisor, Dr. Anitha, for her invaluable suggestions and unwavering guidance throughout this research. I am grateful to my co-advisor, Mr. Teshager Kassa, for his assistance and commitment throughout this journey.

Special acknowledgment is due to Hawassa memeria police commission and one stop center staffs for invaluable assistance with the criminal data and conducting interviews, as well as to the entire staff at the One Stop Center.

Lastly, I want to convey my deepest gratitude to my family members, especially my husband Tewodros, my sister Bezaye, my brother-in-law Nuno, my parents, and my friends Tigist and Tayech, who supported me at every juncture of my thesis. To my son Kalye, your presence provided me with immense strength whenever I encountered challenges.

DECLARATION

I affirm that this thesis is solely my creation and has not been submitted for any degree at any other academic institution. I have duly acknowledged and referenced all sources utilized in this thesis project.

Name: VERONICA TESFAYE BELAYNEH

Date: _____

Signature: _____

ACRONYMS/ABBREVIATIONS

CSV	Comma Separated Values
DT	Decision Tree
EDA	Exploratory Data Analysis
FN	False Negative
FP	False Positive
FPR	False Positive Rate
IoT	Internet of Things
KNN	K-Nearest Neighbor
MAE	Mean Absolute Error
ML	Machine Learning
NB	Naïve Bayes
OLS	Ordinary Least Squares
RAE	Relative Absolute Error
RF	Random Forest
RMSE	Root Mean Squared Error
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Crime is a widespread issue globally and a serious concern affecting the lives of women. This study aims to employ ensemble learning methods to predict crimes against women in different areas of Hawassa City. The proposed research is based on analyzing crime patterns in previously referenced years. By combining predictions from multiple models, ensemble learning aims to improve the overall performance of the model. The results of the study indicate that the ensemble of machine learning models, particularly the proposed classification models, significantly improves crime prediction compared to traditional methods, as evidenced by improved performance metrics.

The study uses a dataset of 3,318 records, with 454 entries from a One Stop Center, and considers 8 attributes: 'Subcity', 'Kebele', 'Year', 'Age', 'Marital status', 'Time of crime', 'Categories', and 'Crime'. The focus is on predicting specific crime types: forced use, kidnapping, snatching, beating, insult, rape, theft, and threats.

The developed model is based on an ensemble technique which is a Machine Learning (ML) based approach. The model has good accuracy, which is 92.87% accuracy. The accuracy results of various machine learning classifiers for predicting crimes against women in Hawassa City are as follows: Random Forest achieved the highest accuracy at 92.87%, followed by Logistic Regression at 91.57%. The Decision Tree classifier also performed well with an accuracy of 91.37%. Both the SVM and Voting classifiers attained an accuracy of 90.86%. AdaBoost had the lowest accuracy among the evaluated models, with a score of 89.86% In addition to the accuracy, based on the importance of each input feature to the final model, Crime categories take the main influencing share by 70.6% from the total input parameters. According to the gathered data, the model demonstrates notable precision and recall rates when employing the Random Forest classifier, especially in effectively identifying specific classes with high precision and recall. The study contributes to changing traditional police stations and investigating activities by incorporating new technology and data recording methods that are more suitable for contemporary scenarios.

Keywords: Random Forest classifier, Ensemble Method, Machine Learning.

Table of Contents

CHAPTER ONE

1. INTRODUCTION

1.1 Background

A crime refers to an illegal action that is subject to punishment by a governing body or authority. Even though both men and women can be a victims of a crime, gender is one of the key factors that significantly increase vulnerability. Based on the United Nations report, out of gender-specific violence on women many aggressors are found to be men and are motivated by gender considerations such as male power and privileges [1] . Many gender-based violence against women are also considered a violation of human rights in many countries around the world. In Ethiopia, the Criminal Code recognizes most violence against women such as rape (Article 620) and different assaults targeting women incapable of protecting themselves (Article 622 and 623) as crimes [2]. Even though such violations are identified as crimes, the extent is increasing more and more, and it is still a serious problem in Ethiopia. Women are entitled to freedom from violence, harassment, and discrimination. Eliminating the obstacles of unsafe environments can empower women to realize their full potential as individuals and valuable contributors to workplaces, communities, and economies [2]. And yet, crimes against women are still a bit of concern worldwide including in Ethiopia. According to [3], it is identified that women aged between 15 and 49 are found to be targets of physical and sexual violence, and 1 in 3 Women have experienced physical or sexual violence at some point in their lives.

In addition to gender-based violations on women stated as crimes in Ethiopia, other crimes target people irrespective of their gender and are widely considered by many as crimes that affect women more than men. These include robberies and different assaults in public spaces. This, in turn, can lead to fear of using such spaces, and cause women to experience higher levels of anxiety, which may limit “access to and use of cities” [4].

The objective of crime prediction, a significant technology within social computing, is to derive valuable insights from extensive criminal records to forecast future criminal activities. It can help police gather crime information and encourage citizens to be vigilant in specific areas.

The rapid growth of big data, the Internet of Things, and other technologies has increased the use of artificial intelligence in predictive and crime prediction models [5]. Hence, there is a necessity to categorize present crime prediction algorithms and thoroughly compare the attributes and circumstances that significantly impact the analysis of such algorithms.

Predicting crimes against women can help government organizations like police departments to mitigate, or women to take precautions. For predicting crime in general or crime against women, machine learning models have been widely studied [6] and [7]. The generic crime prediction does not identify gender-specific crimes, while the latter focuses specifically on gender-based crimes against women. For example, [8]presents the use of logistic regression to predict whether a given place at any given time is safe for a woman to go to or not. There are many additional examples of research focusing on predicting safety for women [7]. Novel technologies have also been used to improve women's safety. For example, an Android app developed by Eranpurwala based on Emergency Alert System (Distress Call) and Safe Routing [9]are a few to mention.

In this proposed research, the aim is to use ensemble learning methods to predict crime against women. Within this study, a reported set of crime data from Hawassa Police sub-city stations and Hawassa University Referral One stop center be used to select dominant crime types on women to predict crime against women. This means crimes categorized as non-gender based but previously seen by many as dangerous to women in Ethiopia will be investigated during the data exploration phase of the study. Within this study ensemble learning methods used in crime prediction means combining baseline models to build a bigger power full model will be used to identify a method that predicts crime against women.

1.2 Problem Statement

Crime against women is a serious problem in our society, and it is getting worse. This happens because the judicial system is ineffective, the legal rules are weak, offenders are not properly identified, and there are no well-organized security measures in place. As a result, crimes against women are not being managed effectively across different areas, making the situation worse.

The increasing incidence of crimes against women poses a significant threat to their safety and well-being, necessitating effective measures to predict and prevent such occurrences. Despite various efforts by law enforcement agencies, there is a lack of precise predictive models that can identify high-risk areas, vulnerable age groups, and the most common types of crimes against women. This study aims to develop a robust predictive model using machine learning techniques to analyze various factors and features associated with crimes against women. By accurately forecasting potential crime hotspots and high-risk demographics, this model will aid in the implementation of targeted preventive measures, ultimately enhancing the safety and security of women.

Research work said that women were abused every 15 to 20 minutes, whether in busy or quiet environments. Times have changed, but attitudes toward women haven't changed. Issues affecting women continue to worsen despite significant progress in many areas. A woman decides whether to go to a certain place or not. For this decision, use machine learning to give certain places a safety level before going out [10].

Utilizing a machine learning approach can significantly enhance crime prediction and analysis. This approach encompasses regression algorithms for prediction and classification techniques to serve investigative objectives. Regression techniques like multilinear regression represent statistical methods within this framework [11].

Research Questions

The intended investigation seeks to address the subsequent research inquiries:

- Which area is the most dangerous?
- Which age group experiences the highest frequency of crimes?
- Which type of crime occurs most frequently against women?
- What are the best variables/features for predicting crime against women?

1.3 Objectives

1.3.1 General Objective

The general objective of this study is to identify the relationship between women's age, demographic area, and crime type to develop a crime prediction model. This model aims to assist the Hawassa City Police Commission in making informed decisions for detecting and preventing crimes against women. By analyzing these factors, the study seeks to enhance the effectiveness of law enforcement strategies, improve resource allocation, and ultimately contribute to the safety and security of women in the community.

1.3.2 Specific Objectives

To achieve the general objective, this study will address the following specific steps:

- Determine the age group that experiences the highest frequency of crimes against women.
- Classify the types of crimes that occur most frequently against women by examining crime records.
- Identify the best variables/features for predicting crimes against women using machine learning techniques.
- Validate the predictive model's accuracy and effectiveness in identifying high-risk. areas and vulnerable age groups, contributing to enhanced safety and security measures for women.
- Identify appropriate machine learning algorithms and select the suitable ones for crime prediction against women.

1.4 Scope and Limitation of the Study

The scope of this research focuses on ages above 12 and uses three years of data from 2013 to 2015 E.C. in Hawassa City. Due to the data being manually stored on paper and not well organized, a significant amount of time is required to convert and structure it for computer analysis. The major limitation of this study is its focus solely on specific types of crimes: forced use, kidnapping, snatching, beating, insult, rape, theft, and threats.

1.5 Significance of the Study

This proposed study aims at identifying the best ensemble learning model to predict the crime against women in Hawassa. Such information can be used by: by predicting the type of crime using the best ensemble learning method will help the police to act immediately (1) distribute resources to tackle, and/or (2) plan mitigations of crime occurrences against women, and Women to take precautions as well.

1.6 Research Methodology

The first step consists of collecting data from Hawassa police station and Hawassa University Referral Hospital one-stop center and then cleaning and exploring the provided dataset. Once this is performed, the second step consists of the Ensemble Learning model development. The last step consists of consolidating the results from both previous steps towards answering the research questions proposed in section 2. To accurately define the research problem, primary data was gathered by conducting interviews with relevant officers employed within the police commission, supplemented by observations. Then based on the information obtained from these attempts, the overall crime prevention processes of the Hawassa city police Commission were described.

In order to define the research problem properly, primary data was collected by interviewing concerned officers who are working in the police commission as well as through observation. Then based on the information obtained from these attempts, the overall crime prevention processes of the Hawassa city police Commission and One stop center in Hawassa university referral hospital was described. Relevant literatures on machine learning techniques and crime were reviewed from books, journals, proceedings and the internet. The potential of machine learning in general and particularly successful machine learning application in crime prevention was assessed and described.

An initial exploratory analysis will be performed to identify missing data, unexpected values, or any patterns such as biases in reported crimes, correlation between variables and their spatial distribution (e.g. spatial auto-correlation). And finally, based on this exploration, Research work

will address the dataset limitations where possible. The two important outcomes of this step of the research will be a descriptive overview of the data and especially important which crime statistics can be used for the model development.

Baseline models: To the more commonly used models, will refer to them as "baseline" models as this will serve as a base comparison against more advanced approaches that I will test. There are dozens of models that can be considered to be "baseline" and therefore I will test the ones which are commonly occurring in other literature, such as, Naive Bayes, Decision Trees, Random Forest, Support Vector Machine and Logistic Regression.

Ensemble learning: Ensemble learning consists of combining the use of multiple machine learning models at the same time, thus potentially minimizing the negatives of each different model. Ensemble models have been shown to often outperform single model approaches and while single algorithms such as random forest use some of these techniques, the use of ensemble models in crime prediction remains an unexplored opportunity.

Model error will be measured using standard error metrics such as classification tasks, metrics such as accuracy, precision, recall, F1-score, and a confusion matrix approach when measuring categorical values of the crime statistic (e.g. risk of crime occurring). As these depend precisely on which data is made available by the Police department, it is not possible to explicitly define which error metric will be used.

1.7 Thesis Organization

This thesis is structured into five chapters, each focusing on different aspects related to crime against women. Chapter 2 reviews existing literature on machine learning and explores concepts associated with ensemble learning and crime against women, giving an overview of machine learning, its functions, and the application areas of ensemble learning technology. It also examines related works and covers the different types of crime against women. Chapter 3 focuses on data preprocessing, explaining the major steps involved, including the materials used and methods applied. This includes a description of the data, statistical analysis of attribute data, data cleaning, transformation techniques, and feature extraction methods utilized to build the

prediction model. Chapter 4 addresses the experimentation process and the interpretation of results, describing the construction of the model using a training dataset and the subsequent validation of the results with the testing dataset. It also delves into the interpretation of the experimental outcomes. Finally, Chapter 5 concludes the thesis by summarizing the achievement of reasonable accuracy in the model and presenting conclusions and recommendations based on research findings. Overall, these five chapters provide extensive coverage of various aspects related to the prediction of crime against women, including a literature review, data preprocessing, experimentation, result interpretation, conclusions, and recommendations.

CHAPTER TWO

2. LITERATURE REVIEW

2.2 Introduction about Machine learning

A branch of Artificial Intelligence known as "Machine Learning" employs statistical models and algorithms to analyze data and generate predictions [12].

The main advantage of machine learning is the ability to learn and improve from data without being explicitly programmed. Here are some key advantages:

- Automation and Efficiency: Machine learning algorithms can automate complex tasks and processes, reducing the need for manual intervention. This leads to increased efficiency and productivity, as machines can analyze and process large amounts of data much faster than humans.
- Handling Complex and large-scale Data: Machine learning excels at handling complex and large datasets that may be difficult for humans to analyze effectively. It can uncover patterns, relationships, and insights that may not be apparent through traditional data analysis methods.
- Adaptability and Generalization: Machine learning models have the ability to adapt and generalize from the data they are trained on. This means they can make accurate predictions or decisions on new, unseen data, as long as it follows similar patterns to the training data. This adaptability makes machine learning applicable to a wide range of problems.
- Continuous learning and improvement: Machine learning models can continuously learn and improve over time as they are exposed to more data. They can update their knowledge and adjust their prediction based on new information, leading to more accurate and up-to-date results.
- Decision-making and prediction: Machine learning models can make data-driven decisions and predictions based on patterns and trends in the data. This can be particularly useful to inform decision-making.

These days, machine learning (ML) has emerged as one of the most potent tools. It employs various algorithms to identify patterns in data, which produces insights and aids in the improvement of judgments and forecasts. When faced with a challenging assignment or issue that requires a big dataset with numerous variables there are situations where the data cannot be explained by the current formula or equation. Using ML techniques in this situation will be quite beneficial. ML systems use various models or algorithms to find patterns in various kinds of data. The data set might contain quantitative, structured, and unstructured data.

Machine learning algorithms fall into two categories: supervised learning, where the prediction model is created using a known input and output data set, and unsupervised learning, where the learner only has access to the input data and must find hidden patterns that link various variables to create groups of related items.

2.2.1 Machine learning techniques

Machine learning tasks can be divided into two main categories supervised learning and unsupervised learning.

Supervised learning

In a supervised setting, the fundamental idea revolves around furnishing the agent with a precise indication of the error (which can be directly compared without specific values) in relation to actual algorithms. This function is facilitated by a training set consisting of pairs of input and expected output [13].

Examples include linear regression, Logistic regression, support vector machines (SVM), decision trees, and Random Forests

One common application of supervised learning across predictive analysis involves utilizing regression or categorical classification methods.

- Classification: The basic idea behind classification jobs is to anticipate a particular category of data (discrete variables). Classification is a subset of machine learning tasks that includes determining an item's group or category based on specific attributes. To achieve this, preassign labels to the data, informing the computer of the class to which

each item belongs. Subsequently, train the computer using diverse techniques until it acquires the capability to accurately recognize each class. In short, the output of a classification task is a model that can be used to identify the class to which a new individual belongs. Class probability estimation, often known as scoring, is a closely related task. When a scoring model is used on an individual [14].

- Regression: The primary focus of regression tasks is numerical value (continuous variable) estimate. A supervised machine learning method called a regression task is used to forecast the values of a target variable by using one or more independent variables as input. The goal is to minimize the sum of squared errors between the observed data and the projected values by fitting a mathematical model to the observed data points. Both linear and non-linear models are used in regression tasks to create our predictive models. Non-linear models do not rely on any of the fundamental assumptions made by linear models, which are that there is a linear relationship between the input and output variables. Finding is the aim of linear regression [14].

Both regression and classification are machine learning algorithms that learn a mapping function from the input data. The errors that learning algorithms commit while they are learning are referred to as bias and variance. The bias indicates how well the model captures the mapping function between inputs and outputs. Conversely, the variance of the model shows how the model's performance changes based on the training set. There is a relationship between a model's variance and bias. It is better to have a model with minimal variance and bias. A biased dataset that inaccurately represents a model's use case leads to low accuracy levels and analytical errors.

Unsupervised Learning

The objective is to uncover intricate processes and underlying patterns within input data without predefined labels. This method operates without supervision, relying solely on the data's characteristics. Absolute error measures become valuable when understanding how elements can be grouped based on their similarity, using techniques like clustering. Clustering stands out as one of the primary methods in unsupervised learning, aiding in the identification of patterns and structures within datasets without explicit guidance from labeled responses [15].

- Clustering: Clustering, a frequently employed machine learning technique, involves categorizing data points into clusters or sets of closely associated data points. This unsupervised method operates without the need for labeled data and serves to reveal patterns or resemblances within a dataset. Clustering finds diverse applications, including customer segmentation, market analysis, image categorization, and document organization, among others. Fundamentally, clustering entails partitioning a set of objects into separate groups, with the aim of ensuring that elements within each group exhibit similarity, while those in different groups show distinct dissimilarities [15].

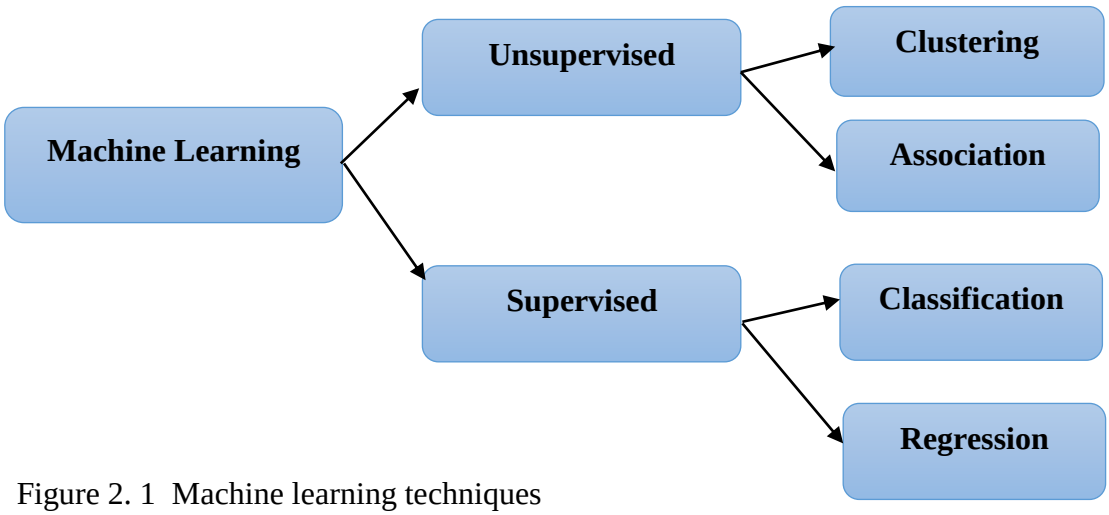

Figure 2. 1 Machine learning techniques

2.2.2 Predictive Modeling

Predictive modeling is a sophisticated methodology that utilizes advanced mathematical and computational techniques to anticipate future events or results. It involves training a model by analyzing a vast amount of example data, which can include various sources such as images, webpages, journal articles, speeches, and social media profiles. In statistical terms, the variables that are manipulated independently are called input variables, while those whose values depend on other factors are termed outcome variables. However, in machine learning, these are referred to as features and goal values or labels, respectively. The primary objective of supervised machine learning or predictive modeling is to develop techniques capable of accurately predicting outcomes or targets [16].

2.2.3 Classification Learning Algorithm

In the realm of machine learning, classification is recognized as a supervised learning method within predictive modeling [17].

- Binary classification: pertains to tasks where data is categorized into two distinct labels, such as "true and false" or "yes and no."
- Multiclass classification: on the other hand, encompasses tasks with more than two classes. Throughout the machine learning and data science literature, numerous algorithms have been proposed for classification, with several widely utilized across various domains [17].

The effectiveness of classification algorithms can be gauged by factors like their computational complexity, accuracy, and scalability. Below are some notably efficient classification algorithms in the field of machine learning [18].

Logistic regression (LR):

Logistic regression stands out as a highly utilized machine learning tool for binary classification tasks [19]. In both social and natural sciences, logistic regression holds significant importance as one of the fundamental analytical techniques. When it comes to classification tasks in natural language processing, logistic regression emerges as the conventional supervised machine learning approach, bearing resemblances to neural networks [20].

Naive Bayes algorithm

The naive Bayes algorithm operates on Bayes' theorem, assuming independence between each pair of features. These classifiers belong to a family of straightforward probabilistic classifiers, applying Bayes' theorem with strong, naive assumptions of independence among features. Since the 1950s, there has been extensive study of naive Bayes. Naive Bayesian networks (NB) are basic Bayesian networks comprising directed acyclic graphs, featuring both unobserved and observed nodes. They strongly assume independence among observed nodes concerning the unobserved nodes [21]and [22] .

Therefore, the independence model (Naive Bayes) is based on estimating as a Formula. Instead, it retains all instances associated with the training data within an n-dimensional space.

Naive Bayes classifier (NB) $\quad p(A/B) = \frac{p(B/A) \cdot p(A)}{p(B)}$

Support Vector Machine

Support Vector Machine (SVM) stands as an effective algorithm utilized primarily for binary classification tasks. It demonstrates computational efficiency, particularly when employing linear kernels, and exhibits proficiency in handling high-dimensional datasets. In the realm of machine learning, another frequently employed technique for both classification and regression tasks is Support Vector Machine (SVM). This state-of-the-art method, along with various supervised learning models, employs related learning algorithms to analyze data for both regression and classification analyses.

Decision Tree

Decision tree (DT) is a widely recognized non-parametric supervised learning technique applicable to both classification and regression tasks. Decision tree learning methods are commonly utilized for creating classifiers, with decision-tree representation being a popular logic technique for this purpose. The literature on machine learning and applied statistics extensively discusses various decision-tree induction techniques [23]. Decision tree learning utilizes a decision tree as a predictive model, wherein observations pertaining to an item are linked to conclusions regarding the item's target value. It serves as one of the predictive modeling approaches employed in statistics.

Random Forest

Although decision trees are commonly used in supervised learning, they can encounter issues like bias and overfitting. However, these challenges can be mitigated by employing the random forest algorithm, where multiple decision trees collaborate to generate more precise predictions, especially when the trees are uncorrelated. In a random forest, several tree predictors work together, with each tree depending on values from a random vector sampled independently and uniformly across all trees. As the quantity of trees within the forest grows, the generalization

error of the forest typically reaches a point of stabilization. Essentially, a random forest is an ensemble learning technique that amalgamates predictions from multiple decision trees to yield more accurate and consistent forecasts. It operates as a supervised learning algorithm suitable for both classification and regression tasks. By incorporating bagging and feature randomness, the random forest algorithm extends the capabilities of the bagging method to create a diverse set of uncorrelated decision trees [24].

2.3 Ensemble Method Algorithms

2.3.1 Ensemble method

Ensemble methods in machine learning entail constructing a group of predictive models, the outputs are combined to generate a unified prediction. The main aim is to improve predictive precision, a concept backed by various studies showing the superiority of ensembles compared to individual models. Although the roots of ensemble methods can be traced back to the 1970s, it wasn't until the 1990s, with the introduction of techniques like bagging and boosting, that these methods gained widespread acceptance. Presently, ensembles are regarded as a foundational approach in machine learning, indispensable for achieving high levels of predictive accuracy, particularly when accuracy is of utmost importance [25]And [26].

In certain datasets, a single algorithm might struggle to produce the most accurate predictive model due to the diverse constraints inherent in different machine learning algorithms. Achieving high accuracy in the modeling phase poses a significant challenge. To overcome this obstacle, various methods exist to enhance overall model accuracy. One effective strategy involves generating multiple sub-models and integrating their outputs into a final model.

Ensemble techniques present a machine learning approach wherein multiple base estimators or models are merged during the modeling process. These ensemble learners provide a solution to the limitations associated with using a single algorithm in machine learning. They aim to improve predictive performance by reducing errors in regression tasks and boosting classification accuracy.

2.3.2 Ensemble Techniques

There are different techniques used in ensemble techniques, but the most commonly used techniques in ensemble method are [26]:

1. Bootstrap Aggregating (Bagging): This technique involves generating base estimators or models using the same algorithm.
2. Adaptive Boosting (AdaBoost): In this method, weak learner algorithms serve as base estimators to create a robust final model. It operates by iteratively applying a weak learner, such as classification rules or decision trees, on different subsets of training data. The classifiers produced by these weak learners are then amalgamated into a unified strong classifier to achieve higher accuracy than that of the individual weak learner's classifiers.

2.3.3 Hyper Parameters

In machine learning algorithms, there are parameters that dictate how input data is transformed into desired outputs, known as model parameters. Conversely, some hyperparameters do not directly learn from data but rather shape the overall structure of the model. Particularly in tree-based machine learning algorithms, these hyperparameters significantly influence the model's architecture. To identify the most optimal model structure, it is crucial to explore the range of possibilities for these hyperparameters. This endeavor, known as hyperparameter tuning, aims to fine-tune the model's performance by optimizing its configuration.

There are three ways to search the optimal hyperparameter values

• Grid search: This method involves creating models for every possible combination of hyperparameter values, assessing each model's performance, and selecting the architecture that yields the best results.

• Random search: In this approach, models are built using randomly chosen hyperparameter values, sampled based on the statistical distribution of each hyperparameter.

• Bayesian optimization: Unlike the first two methods which optimize models independently using various hyperparameter values, Bayesian optimization leverages the outcomes of previous iterations to enhance the selection of hyperparameters for the next experiment.

1. Bagging (Bootstrap aggregating)

The Bootstrap Aggregating method is based on the idea of "bootstrap "which is a powerful statistical method for estimating a quantity from a data sample. In this method, if the number of samples is small in number, the mean will have an error in it. To improve the mean, we use the bootstrap procedure:

- Generate numerous randomized sub-samples from the dataset, allowing for replacement, meaning any value can be chosen multiple times within each sub-sample.
- Compute the mean for each of these sub-samples individually.
- Determine the average of all the means obtained from the sub-samples, treating this value as the mean representative of the original dataset.

2. Boosting

Boosting stands out as a powerful ensemble learning approach, leveraging the collective strength of multiple weak learners to forge a robust predictive model. This method operates sequentially, with each weak learner honing in on the instances misclassified by its predecessors. AdaBoost, for instance, strategically assigns greater emphasis to misclassified data points in successive iterations. Gradient Boosting, on the other hand, orchestrates a meticulous dance of decision trees, each refining the predictive accuracy by targeting the residual errors of its predecessors. Notably, XGBoost raises the bar further with its innovative blend of regularization techniques and parallel processing capabilities, propelling boosting to the forefront of machine learning performance.

3. Voting

This approach is a simple but powerful way to combine predictions from different models. For classification tasks, it works like a democracy called majority voting. Imagine everyone in a group voting on an issue. The decision is made based on what most people agree on. Similarly, in this strategy, the final prediction is the one that most base learners agree on. It's a

straightforward way to make a decision that represents the collective opinion of all the models involved.

This technique offers versatility through its two principal methods:

- Hard voting is like when most people agree on something, so that's the final decision. It's like saying "yes" or "no" based on what most people say.
- Soft voting is more about considering how sure each person is about their opinion. It looks at the chances or probabilities each person gives. Then, it combines all these opinions to make a decision. So, it's more like weighing everyone's opinion based on how confident they are.

Both ways have their uses, and they help to make the technique adaptable to different situations.

Voting involves combining different opinions to make decisions and in voting, gathering people's opinions to make decisions that represent what most people want. So, both methods show how working together can lead to better results and harmony.

Numerous widely used ensemble algorithms adopt this methodology, including:

- Bagged Decision Trees (canonical bagging),
- Random Forest algorithm that makes a small tweak to Bagging and results in every powerful classifier
- Extra Tree

2.3.4 Crime against women Prediction overview

Over the past few years, machine learning has gained a lot of popularity, moving from being used to predict future company investments. A growing number of studies on the application of machine learning to forecast the rate of crime against women have been published throughout the years.

2.3.5 Crime against Women

"Crime against women" encapsulates a spectrum of egregious acts, encompassing not only physical assault but also psychological and sexual abuse, inflicted by both strangers and,

regrettably, at times, by those closest to them. The silent anguish endured by victims often goes unnoticed, lacking the public acknowledgment it deserves. However, the tide began to turn in the 1980s as grassroots and global women's movements rallied to spotlight the pervasive nature of this injustice. Slowly but surely, crime against women has been recognized not only as a grave threat to their health and well-being but also as a fundamental violation of human rights. For over three decades, women's advocacy groups have tirelessly championed action and heightened awareness surrounding the sexual, psychological, and physical exploitation of women, igniting a fervent call for change [27].

Crimes against women often stem from a complex interplay of societal attitudes, entrenched gender disparities, power dynamics, cultural norms, educational deficiencies, economic pressures, and psychological influences. Addressing these root causes demands a multifaceted approach centered on promoting gender equality, enhancing education and awareness, and bolstering law enforcement effectiveness. Concurrently, empowering women to take precautions can serve as a vital component of prevention strategies. In the realm of predictive analytics, ensemble models stand out for their ability to integrate diverse algorithms such as decision trees, random forests, and gradient boosting. By leveraging these techniques, each model within the ensemble can be trained on distinct subsets of data or with varying features, enriching the predictive capacity of the overall model. When applied to the prediction of crimes against women, an ensemble model would require a comprehensive dataset incorporating pertinent features such as demographic profiles, historical crime data, and nuanced information specific to crimes targeting women, including victim profiles. By assimilating these multifaceted elements, the model can discern intricate patterns and correlations, thereby furnishing insights crucial for preemptive interventions and targeted prevention efforts.

The ensemble model would be trained on the labeled dataset, where past instances of crime against women are identified. The model learns from this data to recognize patterns and relationships between the features and the occurrence of crime against women. Once trained, the model can then be used to predict the likelihood of feature crimes against women based on new input data.

2.3.6 Feature importance

Feature importance entails a set of methods used to assign rankings to input features within a predictive model, indicating their relative significance in prediction outcomes. These scores are applicable in both regression, where numerical values are predicted, and classification, where class labels are predicted. They serve as valuable metrics across various scenarios within predictive modeling tasks:

- Feature importance scores offer insights into the dataset: They reveal the relative significance of features regarding the target variable, identifying both the most relevant and the least relevant features.
- Feature importance scores offer insights into the model: These scores, typically derived from a fitted predictive model, shed light on the specific model's behavior, indicating which features carry the most and least weight in predictions.
- Feature importance aids in model enhancement: By leveraging importance scores, one can decide which features to retain (highest scores) or discard (lowest scores), thereby refining the predictive model.

2.4 Related Work

V.Shivaprasad et al. [27] More The goal of the study is to forecast whether a woman can safely visit a specific location at any given moment. A machine learning project predicts female safety with 86% accuracy using logistic regression and the dataset consists of 2580 rows and 10 columns. The paper is good even for its preventive mechanism for women to protect themselves before going out. However the paper didn't focus on the crime that will happen to women only knowing the places and the research work didn't put any crime type occurrence and also compared three algorithms the K-Nearest Neighbors, Logistic Regression, Machine Learning, Naïve Bayes but no reason is put why only this algorithm only chosen. Process effectiveness is low; the best one has (the one they left) more performance. The new proposed work will not compare the model instead using different models together to get good performance and the existing work needs improvement.

Tanya Singh et al. [28]The study "Analysis, Forecasting and Prediction of Crime Against Women Using Machine Learning Techniques" explores the application of machine learning to address the pervasive issue of crime against women in India. It employs the Random Forest algorithm for crime prediction and the ARIMA model for forecasting based on data sourced from the National Crime Records Bureau. The research achieved an accuracy of 80% in predicting crimes such as rape and assault, demonstrating effectiveness in identifying crime patterns and predicting future occurrences. By leveraging these models, the study aims to support law enforcement in preemptive interventions and enhance women's safety through proactive crime prevention strategies.

Betelhem Zewdu Wubineh [29].The paper "Crime Analysis and Prediction Using a Machine-Learning Approach" explores the application of machine learning techniques to predict crime types based on historical data, aiming to assist law enforcement agencies in improving their crime investigation efforts. T The study utilized a dataset obtained from the criminal records of the Hossana Police Commission, consisting of 1600 records and 13 attributes the study evaluates three algorithms: Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KNN), finding that RF achieved the highest accuracy of 86.07%. DT and KNN followed with accuracies of 84% and 81%, respectively. While RF outperformed the others, the study acknowledges limitations such as dataset specificity, incorrect classifications, and the narrow range of algorithms tested. The experiments demonstrated RF's promise in crime classification, though the effectiveness of algorithms may vary with different datasets. The paper highlights the importance of continuous evaluation and adaptation of predictive models to ensure their effectiveness, emphasizing the potential of machine learning to enhance law enforcement's crime prevention and investigation efforts.

Vinay Narayan Bhat et al. [30]The paper by Vinay Narayan Bhat, V Santhosh Kumar, and Prof. Saravanan C presents a comprehensive analysis and prediction framework for crimes against women in India using machine learning algorithms. Highlighting the significant rise in such crimes, the study utilizes data from the National Crime Records Bureau (NCRB) to predict crime rates across different states and union territories. Employing linear regression and random forest algorithms, the authors find that linear regression achieves superior prediction accuracy compared to random forest, reaching an impressive 83%. The methodology involves rigorous

data preprocessing to clean and refine the dataset, ensuring robust model performance. Implemented within a Flask-based web portal with MongoDB integration, the system provides accessible crime analytics and predictive insights, supporting proactive law enforcement strategies and policy interventions to mitigate crimes against women effectively in India.

Purvi Prasad et al.[31]In their comprehensive study on crime against women in India utilized Huber regression and time series algorithms to analyze and predict crime patterns across different states and union territories. Their research aimed to address the increasing rate of crimes against women using data mining techniques. The Huber regression method was applied to analyze various types of crimes, including rape, kidnapping, dowry deaths, and assaults on women, across different regions. They converted time-series data into supervised learning formats to predict future crime occurrences. Their findings indicated that Huber regression provided accurate predictions close to actual crime numbers, enabling effective visualization of crime trends. This approach assists law enforcement and policymakers in implementing targeted interventions to improve women's safety, making it a crucial tool in combating crime against women in India.

S. Lavanyaa et al [32].In their study presented at the International Conference on Advances in Signal Processing, Power, Embedded, Soft Computing, Communication and Control Systems (ICSPECS-2019), focused on crime against women (CAW) analysis and prediction within Tamil Nadu Police using data mining techniques. Their research detailed a structured approach involving data preprocessing, rule induction, decision tree modeling, and evaluation phases to classify and predict crimes such as dowry deaths, women harassment, and child abuse. They highlighted the importance of tools like WEKA and MATLAB, with Naïve Bayes and Apriori algorithms showing significant accuracy in crime prediction. Their findings underscored the effectiveness of data mining in enhancing law enforcement strategies and reducing crime rates against women in urban settings but doesn’t specify exact accuracy results.

W.S.V. Lakshan et al. [33]The paper aims to present an enhanced crime prediction algorithm based on ensemble classification techniques, taking into account factors that affect the learning model performance. It defines several factors that influence crime prediction and analyzes their correlation with the prediction label. The proposed algorithm utilizes the Random Forest model

as the base model and Logistic Regression and K-Nearest Neighbor algorithms as sub-models. It develops a final classification using Graphical User Interface and REST API methods to predict the possibilities of crime occurrences at specific times and locations. The study contributes to changing traditional police stations and investigating activities by incorporating new technology and data recording methods that are more suitable for contemporary scenarios and the accuracy of the crime prediction algorithm was 89%. The paper's proposed methods enable the implementation of better strategic and tactical approaches to minimize crimes with less risk.

Table 2. 1 Summary of Related Work

Title	Authors	Methodology	Drawbacks	Results of paper
Women Safety Prediction using Logistic Regression Model	V.Sushma Swaraj et al. 2020	Logistic Regression, Naive Bayes, K-Nearest Neighbours used for classification. Numpy, Pandas libraries utilized for data analysis and visualization	Lack of comparison with other safety prediction models. Absence of detailed discussion on real-time implementation challenges	Logistic regression model predicted women's safety with 85% accuracy. Precision value obtained from experimental results was 88%. Machine learning model achieved higher accuracy with 70% training data.
Analysis, forecasting and prediction of crime against women using machine learning techniques	Tanya Singh et al .2023	Random Forest Algorithm and ARIMA model	Limited accuracy in current crime prediction models for women's safety	Random Forest and ARIMA 80% effectively forecast crimes against women

Crime analysis and prediction using machine-learning approach in the case of Hossana Police Commission	Betelhem Zewdu Wubineh.2024	Supervised learning method for crime classification using decision tree, random forest, K-nearest neighbor	Need for parameter tuning to enhance algorithm performance	Random forest achieved 86.07% accuracy
Analysis and Prediction of Crime against women in India using machine learning algorithms	Vinay Narayan Bhat et al. 2021	Linear regression and random forest algorithms for crime rate prediction	Undefined features dropped due to lack of relevance.	Linear Regression had 83% accuracy

CHAPTER THREE

3. METHODOLOGY FOR CRIME AGAINST WOMEN PREDICTION

The section of the research paper that is devoted to materials and methods provides a detailed explanation of the materials used in conducting the study and the experimental procedures followed. This section is crucial because it enables other researchers to replicate and validate the study's findings. Additionally, the materials and methods section should include information about the data collection and analysis process, such as the type of data collected, specific measurements taken, and any test or statistical methods used for analysis. Any software programs used should also be described, including their version and relevant settings. Finally, a detailed description of the dataset, including data cleaning, data pre-processing, and Data Exploratory, should be provided.

3.1 Data understanding

This section covers the responsibilities of data preprocessing and interpretation in machine learning. Additionally, various strategies have been employed to address the issues mentioned in Chapter One. This includes a comprehensive description of the dataset, as well as an explanation of the data cleansing, transformation, and integration processes. The data contains information on crimes, including those that have not yet received a final decision by the court, and those that have been investigated and resulted in a final decision. The data also includes information on individuals under the age of 18 who have been involved in crimes.

3.2 Data Source and Collection

Data collection is a systematic practice that involves gathering relevant information to build a consistent and complete dataset for a specific business purpose. This process is crucial for decision-making, answering research questions, or strategic planning. It usually entails manually collecting data from each sub-city to create a comprehensive dataset. In summary, data collection is the initial and necessary stage of any activity that involves data.

Data collection is the initial phase of the decision-making process guided by machine learning. In machine learning projects, data collection precedes such stages as data cleaning and preprocessing, model training and testing, and making decisions based on model output.

The data for this research has been collected from Hawassa each 7 sub City police memereya from the daily crime registration book and from Hawassa University Referral Hospital stop center that females less than 18 years got Gender-based violence there is crime registration in one stop center from 2013 to 2015 some of the data does include crimes which have final decision another one is it's collected from daily crime register book also some of the women attacked when she is one the way she doesn't know even who they are(attacker) and only she came to the police station and register what she happens on her. The data collected includes all of the crime.

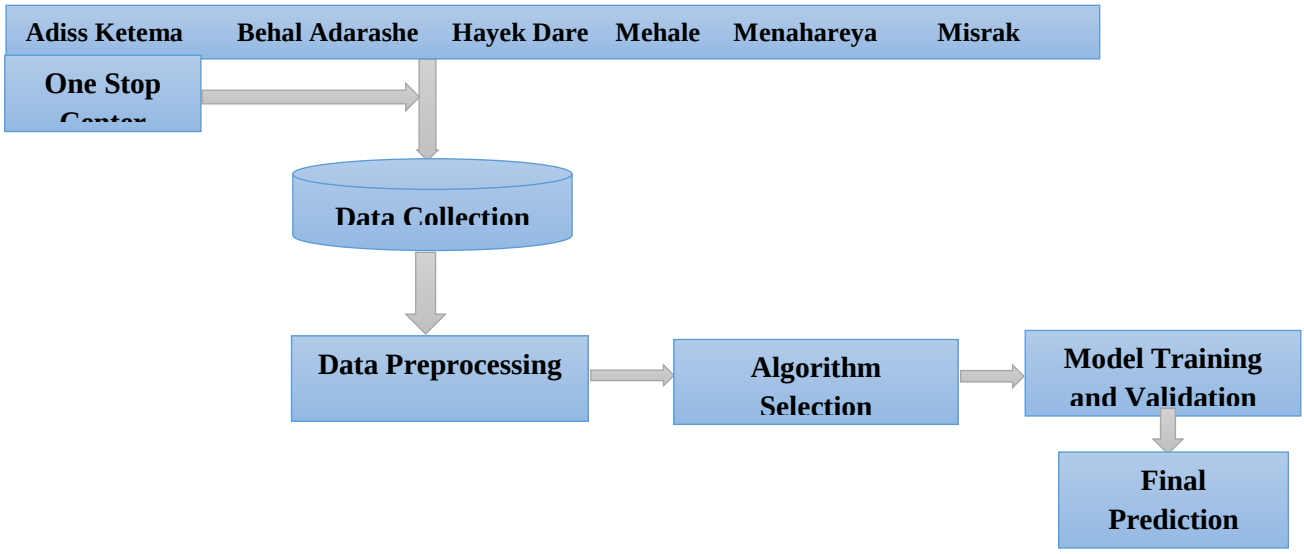

Figure 3. 1 Research methods

3.3 Description and Quality of Data

The number of crime records available in the Hawassa sub-city from 2013 -2015 E.C 7,583 data. From those records around 3318 crime against records were registered on daily crime books and also in crime investigation reports. The only crime data used for this investigation was a crime against women.

There are a lot of attributes in the police station data, but criminal issues require privacy and are sensitive, such as names and specific addresses (phone numbers). Therefore, those details are excluded from the paper, which contains 12 attributes after basic information is filtered out of those attributes, only women attackers are filtered out, and the remaining attributes are based on additional attributes related to the crime against women category in the literature sub-city consulting with domain experts, only pertinent qualities were chosen to be used in feature selection for task execution.

Table 3. 1 Crimes Against Women Data collection description and Number of records

Name of Data source center (Hawassa city)	Year	Number of crime reports	Number of Features	Data type
Tabor	2013-2015	851	8	Number and Text
Hayk Dar	2013-2015	525	8	Number and Text
Bahel Adarash	2013-2015	482	8	Number and Text
Menehariya	2013-2015	438	8	Number and Text
Adiss Ketema	2013-2015	434	8	Number and Text
Mehale	2013-2015	295	8	Number and Text
Misrak	2013-2015	293	8	Number and Text
Total Number of Records		3318		

3.4 Data Collection Methods

Data collection is the process of collecting and evaluating information or data from multiple sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities. It is an essential phase in all types of research, analysis, and decision-making [39].In the data collection scenario, there are two methods Primary and secondary method. Let us see both of them:

1. Primary data

Primary data are those that have been gathered via first-hand experience. Primary data is more trustworthy, genuine, and objective because it hasn't been released yet. Primary data is more valid than secondary data because it has not been modified or tampered with by humans.

Gathering primary data entails getting original information straight from the source or by speaking with respondents face-to-face. Using this approach, researchers can get first-hand data that is well-suited to their goals. Experiment, Survey, Questionnaires, Interview [39].

2. Secondary data

Data collected from a source that has already been published in any form is called secondary data. The review of literature in any research is based on secondary data. It is collected by someone else for some other purpose. Books, Records, Newspapers, Internet articles, and Research articles by other researchers (journals) [39].

3.5 Sampling Techniques

The method of collecting data from a population, regarding a sample on a group of items, and examining it to draw out some conclusion.

Random Sampling

As the name suggests, in this method of sampling, the data is collected at random. This implies that each element in the universe has an equitable opportunity to be chosen for investigation purposes. In other words, each item has an equal probability of being in the sample, which makes the method impartial.

3.6 Data pre-processing

The data pre-processing step is a crucial part of preparing the dataset for training with the selected algorithm. A total of 8 features are collected from records, which are often challenging to obtain due to their sensitive nature and the need for manual data entry. Some records have missing values, which affects the features selected based on the record type. Another issue is

the variation in data types, as the collected data includes both numeric and non-numeric categorical data types including Subcity, Kebele, Marital status, Time, Category, and Crime type. To handle non-numeric data types, we use categorical encoder libraries such as 'import category encoders as ce' in Python and get dummies. These libraries offer a collection of transformers, akin to those in scikit-learn, designed to convert categorical variables into numerical equivalents using various methods. Some basic data preprocessing techniques commonly used in data analysis and machine learning [40]and [41]:

1. Data Cleansing: Data cleaning stands out as a pivotal element of machine learning. A meticulously cleaned dataset heightens the likelihood of achieving favorable outcomes even with basic techniques, which can prove particularly advantageous, especially in scenarios involving extensive datasets and computational complexities. Various types of data call for distinct cleaning procedures, yet adopting a systematic approach is consistently a prudent initial step. Given that raw data often contains noise, inaccuracies, and inconsistencies, data cleansing emerges as a vital prerequisite to ensuring the accuracy and reliability of insights derived from it. Before delving into the code, data cleaning is imperative.
    - Eliminating Repetitions: Identify and discard duplicate entries within the dataset.
    - Rectifying Mistakes: Identify and rectify any discrepancies or inaccuracies present in the data.
    - Precision: Clean data guarantees that analyses and machine learning models are founded on precise and trustworthy data.
    - Uniformity: Incoherent data may lead to inaccuracies, particularly when dealing with categorical variables, date representations, or measurement units.
    - Wholeness: Absent data may impede analysis and modeling efforts. Addressing missing values stands as a vital aspect of data refinement.
2. Dealing with Missing Data:

Identifying missing values in a dataset poses a common obstacle in practical data analysis scenarios. This problem typically emerges from various causes such as human oversight, system errors, or issues during data gathering.

- Approaches for managing missing data encompass:
- Eliminating rows or columns containing missing values.
- Fill in missing values using the mean, median, mode, or alternative strategies.

3. Categorical Variable Encoding:

Translate categorical variables into numerical formats compatible with machine learning algorithms. Methods encompass one-hot encoding, label encoding, or ordinal encoding, chosen based on the characteristics of the categorical data.

4. Feature Scaling:

Normalize numerical features to a comparable range to avoid certain features overshadowing others during model training.

Widely used scaling methods comprise min-max scaling (normalized scaling) and standardization (adjusting to zero mean and unit variance).

Normalization, a widely used method for feature scaling, usually entails adjusting features to fit within a defined range, commonly from 0 to 1. This is accomplished by subtracting the minimum feature value and then dividing by the feature's range, which is the difference between its maximum and minimum values.

Normalization is particularly useful when the features have different units or scales, and it helps prevent features with larger magnitudes from dominating those with smaller magnitudes during the modeling process.

Normalization is particularly useful when the variables have different units or scales, as it helps ensure that all variables contribute equally to the analysis and modeling process. This is important because many machine learning algorithms are sensitive to the scale of the features, and features with larger magnitudes can dominate those with smaller magnitudes if not properly scaled.

5. Managing Outliers:

Detect outliers within the dataset using statistical approaches or visualization methods.

- Strategies for managing outliers encompass:
- Eliminating outliers if they stem from errors or anomalies.

6. Dataset Partitioning:
   - Dividing the dataset into training, validation, and testing subsets to facilitate model training, assessment, and validation, correspondingly.
   - Typical partitioning methods involve the 70/30 or 80/20 train-test split, or utilizing cross-validation for enhanced evaluation reliability.

These are just some of the fundamental data preprocessing techniques used to prepare data for analysis and modeling. Depending on the specific dataset and problem at hand, additional preprocessing steps may be necessary.

3.6.1 Python Libraries for Data pre-processing

Python offers several powerful libraries for data preprocessing. In this article, we'll primarily concentrate on leveraging two popular ones are Pandas: An adaptable library for manipulating and analyzing data and the second NumPy: A foundational package for numerical operations in Python. It is frequently employed alongside Pandas for data cleaning duties

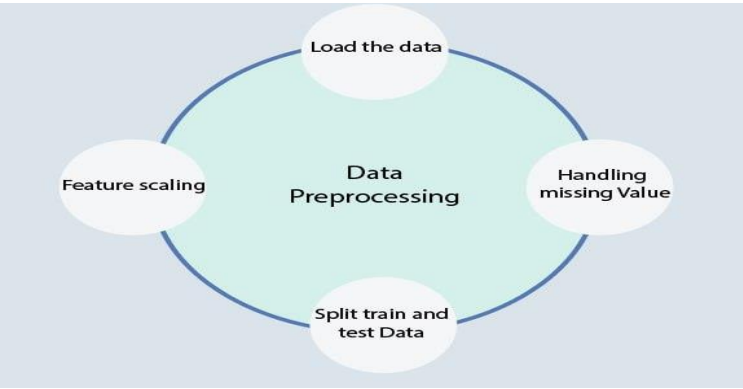

Figure 3. 2 Data pre-processing [42]

3.7 Data Description

After the initial data is collected, the next step is describing the data set. The crime data set has 8 attributes of text and number.

Table 3. 2 Description of Data

No	Feature	Data Type	Description
1	Sub-city	Nominal	The name of the city where the crime occurred.
2	Kebele	Nominal	Specific places where crime happen
3	Year	Number	Year of crime report
4	Age	Number	Age of women, who were affected by the crime.
5	Marital status	Nominal	Marital status women, who were affected by the crime
6	Time of Crime happen	Nominal	The time at which the crime occurred.
7	Categories	Nominal	The general category of the specific crime.
8	Crime	Nominal	Specific crimes occurred.

3.8 Data Exploratory

The process of examining and analyzing record sets to recognize patterns, find outliers, and determine the relationships between variables is known as exploratory data analysis or EDA. EDA is typically done as a first step before more formal statistical studies or modeling is done.

Types of Exploratory Data Analysis:

1. Univariate Non-graphical
2. Multivariate Non-graphical
3. Univariate graphical
4. Multivariate graphical

In this paper from four types of EDA use two of them listed below;

1. Univariate analysis within Exploratory Data Analysis (EDA) centers on the examination of singular variables independently. This analytical approach entails scrutinizing the distribution, summary statistics, and inherent characteristics of each variable within the dataset in isolation, without considering its interactions with other variables.
    - Summary Statistics: Calculation of descriptive statistics such as mean, median, mode, range, variance, and standard deviation to summarize the central tendency, dispersion, and shape of the variable's distribution. data.describe () in python. To analyze the number of crimes, use this type of exploratory data analysis (EDA).
2. Multivariate analysis in Exploratory Data Analysis (EDA) involves examining relationships and patterns between multiple variables simultaneously. Multivariate analysis encompasses various techniques, including:
    - Feature Importance Analysis (Multivariate) Example: Using a Random Forest model to determine feature importance in predicting a target variable. sns.barplot(x=feature_scores, y=feature_scores.index)
    - Correlation Analysis: Calculation of correlation coefficients to measure the strength and direction of linear relationships between pairs of continuous variables. Correlation analysis helps identify associations and dependencies between variables. corr_matrix = data.corr() use this one for correlation heatmap matrix to see the relationship between features.

3.9 Methodology

For this study, the researchers will collaborate with the Hawassa City Police Department and obtain data on reported crimes and the One-stop center. The main objective is to predict crimes against women in Hawassa using this data.

The choice of splitting a dataset into training and testing sets, such as 80/20 or 70/30, is a common practice in machine learning for model evaluation, and there's no one size fits all rule. The decision depends on various factors and it's often influenced by statistical and practical considerations.

It's important to note that these splits are common starting points, and the best choice may depend on factors such as the size of your dataset, the complexity of your model, and the nature of the problem.

The model development process will follow a standard machine learning approach. The research paper tries to address the splitting method by using experiments and getting different results by three methods 80/20, 70/30. To ensure a fair comparison, the study employs an across-validation approach. This involves the training set into several subsets, which will be used to train and test various model hyperparameters and performance metrics. By doing so, the paper can identify the most effective version of the model to address the crime prediction task.

Additionally, the researchers will explore the impact of using different input variables on the performance of the models. This analysis will help determine which variables contribute most significantly to crime prediction.

Overall, this study aims to develop a robust crime prediction model for crimes against women in Hawassa. By employing an ensemble approach investigating the impact of different input variables, the researchers intend to create an effective tool for crime prevention and intervention in the region.

3.10 Modeling

The system model used in this paper to create a model for crime against women prediction using an ensemble model shown below the figure overall steps of the model. Based on all the information collected from the Hawassa city police station and one-stop center.

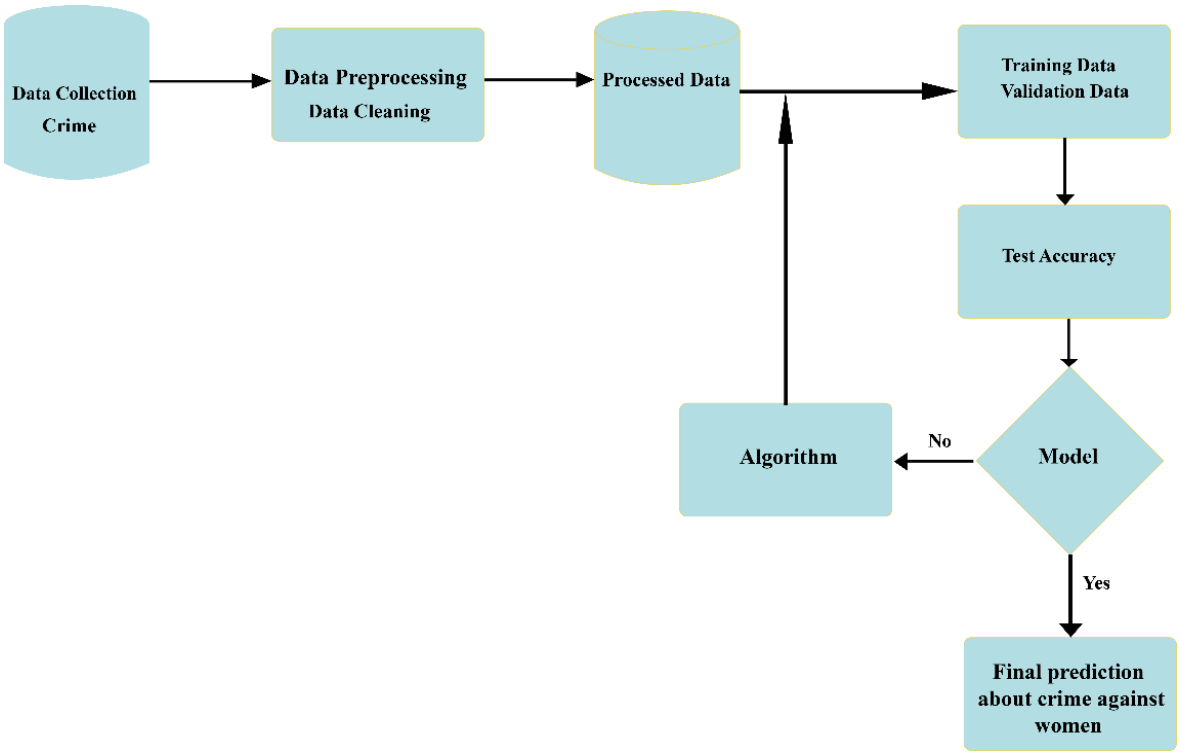

Figure 3. 3 Architecture of the model

3.11 Feature selection

Many of us have faced the task of choosing the most important characteristics from a dataset and eliminating less important or irrelevant features that have minimal impact on our decision-making, aiming to enhance model accuracy.

A variety of methods for feature selection are utilized for diverse objectives:

- By eliminating irrelevant features, it decreases the complexity of the model.
- Assists machine learning algorithms in quicker model training.
- Diminishing dimensionality aids in preventing overfitting.

Feature selection methods for ensemble techniques strive to pinpoint and preserve the most pertinent features while discarding redundant or less informative ones. Several commonly employed feature selection techniques for ensemble methods include:

- Feature Importance Ranking: Ensemble techniques like Random Forest offer inherent feature importance rankings, assessing how frequently features are utilized for splitting and the magnitude of these splits. Features with higher importance scores are deemed more significant and retained for model training, while less crucial features might be omitted.

3.12 Model Evaluation Metrics

The initial dataset undergoes training with an ensemble algorithm to generate the final model in the machine learning process. Before evaluating the model through various methods, k-fold cross-validation is employed to ensure its stability. This involves randomly shuffling the data and dividing it into k subsets. One subset is used for testing while the rest (k-1 subsets) are for training. The training set, typically comprising 60%-80% of the total data, is utilized to build the actual model that will handle new data. Following modeling, the model's performance is assessed using the classification report. In this context, "average weight" and "macro average" are terms referring to different ways of computing performance metrics for multi-classification problems, as utilized in this proposed model [44]:

1. Average Weight (also known as “weighted average”): Weighted average, also referred to as average weight, computes performance metrics by taking into account the support (number of samples) for each class. This approach addresses the imbalance in classification distribution by weighting each class's metrics based on its sample size, thereby prioritizing

classes with larger representation. It is particularly beneficial for imbalanced datasets as it assigns greater importance to classes with more instances.

2. Macro Average: Macro average evaluates a model's performance metrics by treating each class equally, regardless of the number of samples in each category. It independently calculates metrics for each class and then averages them. In essence, the Macro Average assigns uniform weight to each class, irrespective of its representation size. This method is valuable for assessing the model's overall performance uniformly across all classes.

Macro averaging and weighted averaging stand out as prevalent approaches utilized for computing metrics like precision, recall, and F1-score in classification tasks, particularly when confronted with challenges such as imbalanced datasets or multi-class classification problems. These techniques are widely adopted due to their effectiveness in addressing the complexities inherent in such scenarios.

- Misclassification: This is the overall error for the classification model which calculates how much of the data are classified wrongly from the total data set.

3.12.1 Confusion Matrix

A confusion matrix is a technique for calculating the performance of ML classification models. It's just a matrix comparing actual categorical values to expected categorical values. By comparing the actual and predicted classifications that visualizes a classifier's accuracy. When we accurately anticipate actual values, we call them a true positive, and when we predict incorrect values as incorrect, we call them a true negative. A false positive occurs when we anticipate a value that does not occur, and a false negative occurs when we do not predict a value that does occur [45].

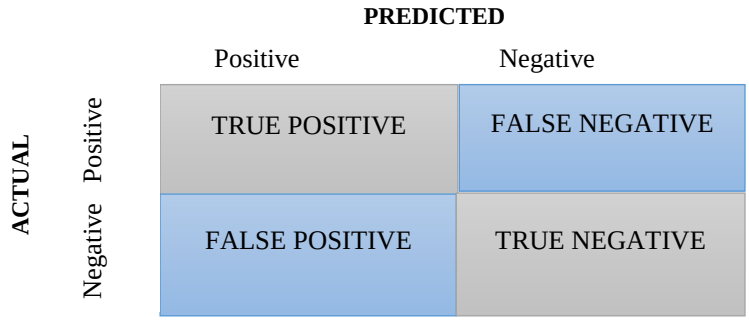

Figure 3. 4 Confusion Matrix

Where:

- TP- True Positive, which predicts actual positive as positive
- TN- True Negative, which predicts an actual negative as negative
- FP- False Positive, which predicts Negative values as positive
- FN- False Negative, which Predicts Positive values as negative.

Note that: TP and TN predict values correctly whereas FN and FP predict values incorrectly.

Based on the confusion matrix values, the performance of the model is evaluated by the following metrics:

1. Accuracy: Accuracy is utilized to determine the proportion of values that are classified correctly. It indicates the number of accurately predicted instances among those classified correctly.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FN} + \text{FP} + \text{TN}) \quad (3.1)$$

2. Precision: It's employed to assess the model's accuracy in correctly identifying positive values. This metric gauges the likelihood of accurately predicting the positive class.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (3.2)$$

3. Recall: It is employed to evaluate the model's capability to forecast positive outcomes. It indicates how accurately we predicted among all the positive instances.

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (3.3)$$

4. F-Measure (F1-Score) F1-Score: is a weighted average score of the true positive

(Recall) and precision.

$$\text{F} - \text{Measure} = (2 * \text{Recall} * \text{Precision})/(\text{Recall} + \text{Precision}) \quad (3.4)$$

3.13 Implementation Tools

3.13.1 Python Programming Language

Python is a versatile, simple-to-learn, and potent programming language. Python is a general-purpose, high-level, interpreter programming language. Initially published in 1991, and created by Guido van Rossum. It enables a more efficient, integrated, and rapid system. It offers straightforward but powerful object-oriented programming techniques along with effective high-level data structures. Python is a great language for scripting and quick application development across a wide range of platforms because of its beautiful syntax, dynamic typing, and interpreted nature.

You can download the large standard library and the Python interpreter for free. New functions and data types written in C, C++, or other languages that can be called from C can be added to the Python interpreter with ease. Python is a good language to use as an extension for apps that may be customized.

Anaconda

Anaconda is a free, cross-platform, open-source distribution of Python and R programming languages tailored for scientific computing needs like data science, machine learning, large-scale data processing, and predictive analytics. Its primary goal is to streamline package management and deployment processes. The package management system Conda oversees the versions of packages, analyzing the existing environment along with any specified version constraints.

Sicikit-learn

Scikit-learn, a Python module, seamlessly integrates traditional machine learning algorithms within the scientific Python ecosystem, which includes numpy, scipy, and matplotlib. Its objective is to furnish straightforward and effective solutions to learning challenges that are accessible to all and can be applied across different scenarios.

Category Encoders

Through a variety of techniques, it transforms categorical variables into numerical values. The current version of scikit-learn provides similar functionalities for ordinal, one-hot, and hashing encoding. Additionally, it boasts compatibility with sklearn, data frames, and column configuration to handle diverse types of input seamlessly.

Matplotlib

Matplotlib, a Python library for 2D plotting, finds applications in various Python environments such as scripts, IPython shells, Jupyter Notebook, Spyder, and web application servers. It facilitates the creation of plots, histograms, power spectra, bar charts, scatterplots, and more.

Testing Environment

This thesis was implemented using the following hardware and software setup.

Hardware Specification:

- System: Intel(R) Core(TM) i7-4600U CPU @ 2.10GHz 2.69 GHz
- Hard disk: 1TB
- RAM: 12GB

Software specification:

- Operating system: Windows 10 Pro,64-bit operating system, x64-based processor
- Coding Language: Python 3.11
- Tools: Jupyter, Notebook, Anaconda, Spyder.

3.14 Feature Importance

It is important to understand the degree of relevance or importance of features that affect the output of target. The score assigned to each feature indicates its significance in influencing the output variable. However, the effectiveness and efficiency of feature importance methods depend on the characteristics of the data, the complexity of the problem, and the type of model being used. Different feature importance methods have their strengths and weaknesses, which should be taken into consideration. Here are a few commonly used methods and some factors to consider when assessing their effectiveness and efficiency [45].

This research work utilizes the Gini importance method to provide unique insights into feature importance in Random Forests. The method measures the randomness of feature importance within a node of a decision tree, which decreases when a specific feature is used to split the data, based on the model's internal calculations [46].

Gini Importance (Tree-based Models): effective for tree-based models like Random Forests and also efficient for tree-based models.

A set of features with associated importance values, typically obtained from a feature importance analysis, possibly from an ensemble model. As explained below in the table and visualized in the graph the information provided:

On the table 4.1, Feature Score Categories (0.70). This feature, labeled as "categories," has an importance value of approximately 0.70. The higher the importance value, the more influential the feature is in making predictions according to the model, Age (0.104) "Age" feature has an importance value of approximately 0.104. This suggests that the model considers age as a relevant factor in its predictions, with a lesser impact compared to the "Categories" feature, Kebele (0.058) "Kebele" feature has an importance value of approximately 0.058. It is another feature that contributes to the model's predictions, though to a lesser extent than "Age." Time of Crime Happen (0.049) The "Timeofcrimehappen" feature has an importance value of approximately 0.049. This implies that the specific time when the crime occurs is a factor considered by the model but with a relatively lower impact compared to other features, Subcity (0.041) "Subcity" feature has an importance value of approximately 0.041. It is another factor the model deems relevant, contributing to predictions but with less influence than the top features, Year (0.025) "Year" feature has an importance value of approximately 0.025. It indicates that the model considers the year in which the crime occurs, but this feature has a comparatively smaller impact on predictions, Marital Status (0.016) "Marital status" feature has an importance value of approximately 0.016. It suggests that marital status is a factor considered by the model, though with relatively low importance.

In summary, these important values give you insights into the contribution of each feature to the model's predictions. Higher importance values indicate more influential features in making

accurate predictions. Understanding feature importance is crucial for interpreting and potentially improving the model's performance.

```
feature_scores = pd.Series(clf.feature_importances_, index=X_train.columns).sort_values(ascending=False)

feature_scores
```

```
categories           0.706114
Age                  0.103594
kebele               0.058836
Timeofcrimehappen    0.049103
Subcity              0.041099
Year                 0.025239
Maritalstatus        0.016015
dtype: float64
```

Using Gini Importance (Tree-based Models) techniques got this result in Table 3.1

Table 3. 3 Feature Score

Features Name	Score (%)
Categories of crime	70.6%
Age	10.4%
Kebele	58%
Time of crime happen	5.1%
Sub city	4.1%
Year	2.5%
Marital status	1.6%

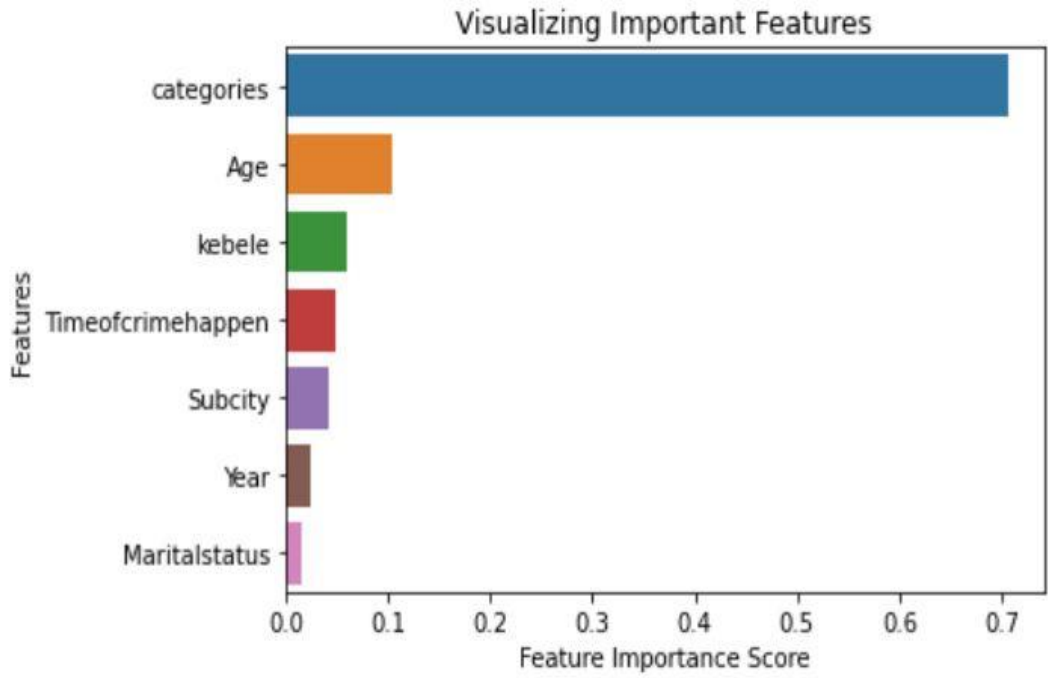

Figure 3. 5 Visualization Important features using Gini importance

3.15 Correlation

There are several reasons why variables in a dataset could be connected.

As an illustration, consider the following:

• One variable may cause or depend on the values of another.

• Two variables may be dependent on a third unknown variable; one variable may have a weak correlation with another.

In data analysis and modeling, it can help to clarify the relationships between variables. The term "correlation" describes the statistical link between two variables.

The correlation matrix you provided shows the correlation coefficients between different variables in the dataset. Here's an interpretation of some key correlations:

Subcity and Kebele: There is a moderate positive correlation (0.35) between Subcity and Kebele. This suggests that there is some degree of association between the subcity and the kebele, indicating that certain kebeles might be more prevalent in specific subcities.

Maritalstatus and catagories: There is a relatively strong positive correlation (0.11) between Maritalstatus and categories. This might indicate that certain marital statuses are associated with specific categories of crimes.

Timeofcrimehappen and Crime: There is a relatively strong negative correlation (-0.24) between the time of crime occurrence and the type of crime committed. This suggests that the timing of crime occurrences might be indicative of the type of crime committed.

Overall: The correlations in the matrix provide insights into the relationships between different variables in your dataset. Further analysis, such as regression modeling or causal inference techniques, may be needed to understand the underlying relationships better.

Techniques that are used in this correlation heatmap is ;

- Matplotlib Heatmap
- Seaborn Heatmap

Both the Matplotlib and Seaborn heatmap methods depict the correlation matrix through a heatmap. Each cell in the heatmap corresponds to the correlation coefficient between two features. The intensity of color within the cells signifies the strength and direction of the correlation: darker shades signify stronger correlations, whether positive or negative, while lighter shades indicate weaker correlations or none at all.

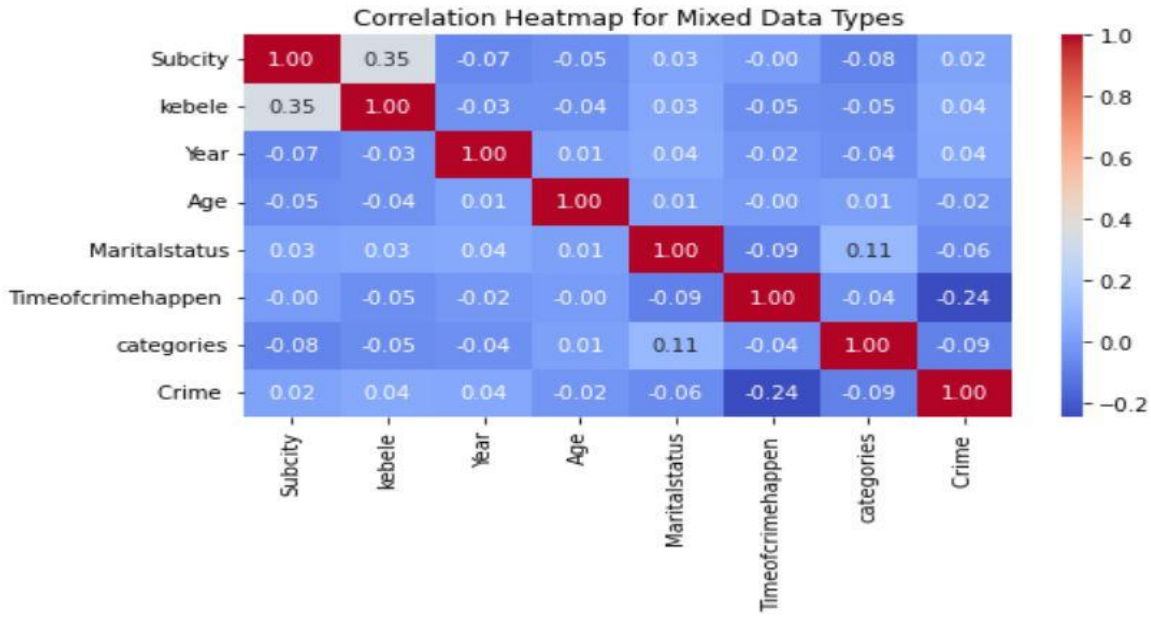

Figure 3. 6 Heatmap Correlation

3.16 Training model

From the supervised ML category, Classification predictive modeling is used to map input variables to discrete output variables. So before training our model need to have load datasets Importing pandas library and assign its ‘pd’, pd.read_csv("veronica.csv"): read_csv function from pandas is used to read the contents of a CSV (Comma-Separated Values) file.

CSV reads the data from specified file ("veronica.csv")and creates a pandas DataFrame,which is a tabular data structure in Python.(data=pd.read.csv).DataFrame is assigned to the variable 'data'.Now 'data' holds the contents of the CSV files,and can performe various operations and analyses on this datasets. data.head() This line uses the head() method to display the first few rows of the data frame. By default, it shows the first 5 rows, providing a quick overview of the dataset's structure and contents.

```
#Importing Dataset
data = pd.read_csv("veronica.csv")
```

```
data.head()
```

	Subcity	kebele	Year	Age	Maritalstatus	Timeofcrimehappen	categories	Crime
0	Tabor	Hogene Wacho	2013	35	Married	Evening	physical	beating
1	Tabor	Hogene Wacho	2013	24	single	Afternoon	pyschological	insult
2	Tabor	Hogene Wacho	2013	31	single	Night	sexual violence	Forced use
3	Tabor	Fara	2013	24	single	Evening	physical	beating
4	Tabor	Hitata	2013	35	single	Afternoon	pyschological	insult

```
data.tail()
```

	Subcity	kebele	Year	Age	Maritalstatus	Timeofcrimehappen	categories	Crime
3313	Mehale	Adiss Abeba	2015	27	single	Evening	pyschological	insult
3314	Mehale	Adiss Abeba	2015	33	Married	Night	physical	beating
3315	Mehale	Nigat Kokeb	2015	29	Married	Night	physical	beating
3316	Mehale	Adiss Abeba	2015	29	Married	Afternoon	psychological	threat
3317	Mehale	Leku	2015	39	Married	Evening	physical	beating

Figure 3. 7 Loading dataset

The choice of splitting a dataset into training and testing sets, such as 80/20 or 70/30.From a total of 3318 final data sets collected 2323 for training and 996 for testing using Python 3.11 ensemble techniques and using RandomForestClassifier

Splitting the Data:

- train_test_split(X, y, random_state=42, test_size=0.3): This line actually perfo rms the split. It takes the feature matrix (X) and target variable (y), and it rand omly shuffles and splits the data into training and testing sets.
- random_state=42: This parameter sets the random seed for reproducibility. If i t uses the same seed, will get the same split each time you run the code.
- test_size=0.3: This parameter specifies that 30% of the data will be used for te sting, and the remaining 70% will be used for training.
- train_test_split(X, y, random_state=42, test_size=0.2): This line actually perfo rms the split. It takes the feature matrix (X) and target variable (y), and it rand omly shuffles and splits the data into training and testing sets.

- random_state=42: This parameter sets the random seed for reproducibility. If i t uses the same seed, will get the same split each time you run the code.
- test_size=0.2: This parameter specifies that 20% of the data will be used for te sting, and the remaining 80% will be used for training.

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
# Normalize the features using StandardScaler
scaler = StandardScaler()
X_train_normalized = scaler.fit_transform(X_train)
X_test_normalized = scaler.transform(X_test)
```

```
# Initialize Random Forest classifier with parameters
rf_classifier = RandomForestClassifier(n_estimators=100, random_state=42)
```

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

```
# Normalize the features using StandardScaler
scaler = StandardScaler()
X_train_normalized = scaler.fit_transform(X_train)
X_test_normalized = scaler.transform(X_test)
```

```
# Initialize Random Forest classifier with parameters
rf_classifier = RandomForestClassifier(n_estimators=100, random_state=42)
```

Figure 3. 8 Training dataset split 80/20 and 70/30

CHAPTER FOUR

4. RESULTS AND DISCUSSION

4.1Analysis of collected data

From the result analysis crime type that mostly happens based on the collected data are beating(36%), insult(13.5%), Forceduse(8.44%), rape(13.2%), theft(19%), Snatch(8.23%), threat(13.2%), kidnapping(0.422%).

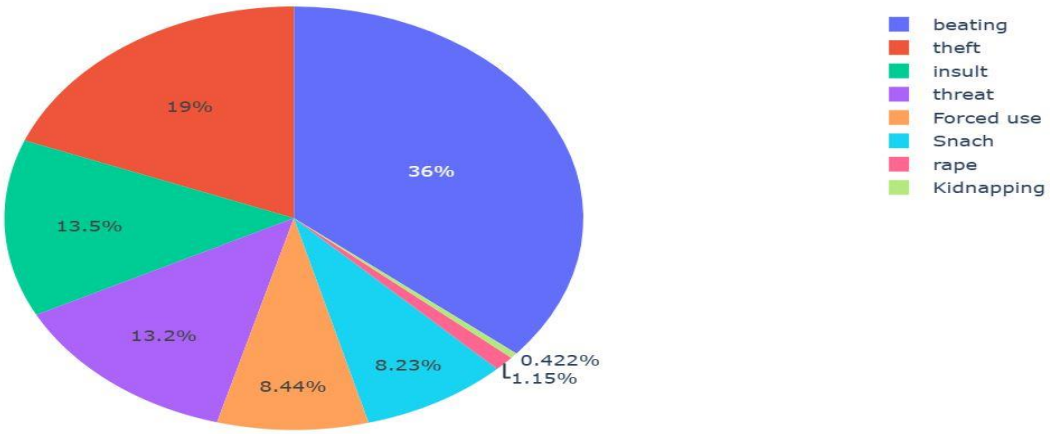

Figure 4. 1 Crime against Women Analysis

The distribution of the target variable classes in the dataset reveals varying frequencies of different types of crimes. Among the recorded incidents, "Beating" stands out as the most prevalent category with 1196 occurrences, highlighting its unfortunate prominence in the dataset. "Theft" follows with 631 cases, indicating a significant frequency of property-related offenses. "Insult" and "Threat" are also notable categories, reported at 449 and 437 instances respectively, suggesting frequent incidents of verbal abuse and intimidation. "Forced use" and "Snatch" are comparatively less frequent but still substantial, recorded at 280 and 273 occurrences respectively. More severe crimes like "Rape" and "Kidnapping" are fortunately less common but remain concerning with 38 and 14 reported cases respectively. This distribution underscores the diverse nature of crimes captured in the dataset, reflecting the range of safety concerns and criminal activities impacting the community.

Table 4. 1 Class name count

Class name	Count
Beating	1196
Theft	631
Insult	449
Threat	437
Forced use	280
Snach	273
Rape	38
Kidnapping	14

Table 4. 2 Performance metrics of classification algorithm result

Performance Metrics of Classification Algorithms:

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Decision Tree	91.3655	91.1347	91.3655	91.2222
Voting	90.8635	90.4593	90.8635	90.6468
Random Forest	92.8715	92.6158	92.8715	92.4781
AdaBoost	89.8594	83.628	89.8594	85.9998
SVM	90.8635	89.9112	90.8635	89.5879
Logistic Regression	91.5663	90.7695	91.5663	90.8181

The results from different machine learning classifiers for crime prediction against women are presented in the table 4.2. Here's a discussion based on the provided metrics (Accuracy, Precision, Recall, and F1-score):

Several machine learning algorithms were evaluated for predicting crimes against women, each demonstrating varying levels of effectiveness across different metrics. The Decision Tree classifier exhibited robust performance with an accuracy of 91.37%, precision of 91.13%, recall of 91.37%, and F1-score of 91.22%. It effectively balanced precision and recall, proving reliable for accurate predictions. The Voting classifier, employing ensemble methods, achieved competitive results with an accuracy of 90.86%, maintaining strong precision and recall rates at 90.46% and 90.86%, respectively, and an F1-score of 90.65%. While slightly lower in accuracy compared to the Decision Tree, it demonstrated effective ensemble prediction capabilities. Random Forest emerged as the top performer, boasting the highest accuracy of 92.87%, precision of 92.62%, recall of 92.87%, and F1-score of 92.48%. Its robustness in handling complex datasets and strong predictive power highlighted its suitability for precise crime prediction against women. AdaBoost showed promising accuracy at 89.86% but exhibited lower precision at 83.63%, indicating potential challenges in correctly identifying crimes against women. SVM maintained consistent performance across metrics with an accuracy of 90.86%, precision of 89.91%, recall of 90.86%, and F1-score of 89.59%, effectively handling diverse data relationships. Logistic Regression achieved an accuracy of 91.57%, with precision and recall scores of 90.77% and 91.57%, respectively, and an F1-score of 90.82%, showcasing its robustness in binary classification tasks. These findings underline the diverse strengths of each algorithm in predicting crimes against women, providing valuable insights for selecting appropriate models based on specific predictive needs.

Discussion:

- **Random Forest** emerges as the top-performing model with the highest accuracy and F1-score, indicating its capability to accurately predict crimes against women.
- **Decision Tree**, **Logistic Regression**, and **SVM** also show strong performance and can be considered reliable alternatives depending on specific requirements such as interpretability or computational efficiency.

- **AdaBoost**, while slightly lower in precision, still provides valuable insights but may require further tuning to enhance precision.
- **Ensemble methods** like Voting and Random Forest validate their effectiveness in combining multiple models for improved predictive performance.

In conclusion, selecting the appropriate machine learning algorithm should consider the specific goals of crime prediction against women, balancing between accuracy, precision, and recall to ensure effective implementation in real-world applications.

The results from various machine learning classifiers for predicting crimes against women reveal distinct performance metrics across models. Random Forest emerges as the top-performing classifier, demonstrating the highest accuracy and F1-score among the models assessed. Its robust performance suggests strong predictive capabilities in identifying and forecasting crimes targeting women. Decision Tree follows closely, exhibiting balanced precision and recall rates, indicating its reliability in accurately classifying instances. Logistic Regression and SVM also perform well, offering competitive accuracy and precision-recall trade-offs suitable for binary classification tasks. AdaBoost, while achieving reasonable accuracy, shows lower precision, suggesting potential challenges in correctly identifying positive instances of crimes against women. Ensemble methods like Voting display effective integration of multiple models but slightly trail behind in accuracy compared to Random Forest. These findings underscore the importance of selecting a model based on specific needs such as accuracy, interpretability, and computational efficiency when developing crime prediction systems aimed at enhancing safety and security for women.

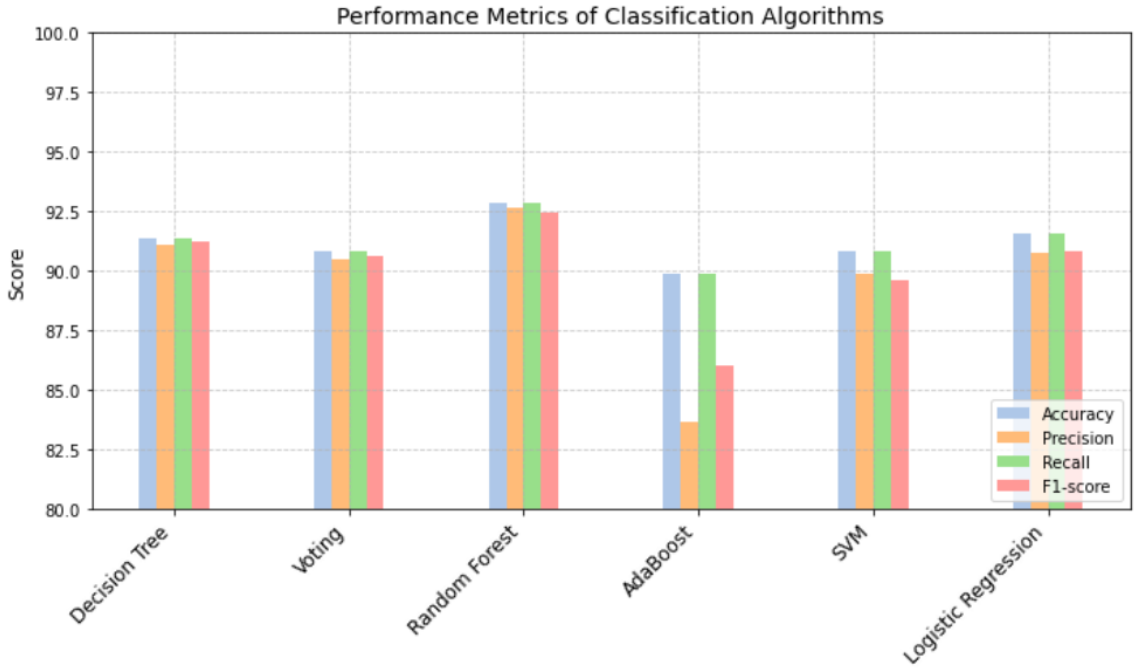

Figure 4. 2 Performance metrics of classification algorithm graph

Table 4. 3 counting crime occurrence based on age group

Age Group	Number of Crimes	Percentage of Total (%)
18-29	1714	51.66
30-44	1361	41.02
45+	166	5
12-17	77	2.32

The Table 4.3 provides a clear breakdown of crime occurrences by age group based on the dataset analyzed. It reveals that the majority of crimes reported involve individuals aged between 18 and 29 years, accounting for 51.66% of the total crimes. The next significant group is individuals aged 30-44 years, comprising 41.02% of the crimes. Older age groups, specifically those aged 45 and above, represent a smaller proportion of the total crimes, with individuals aged 45+ accounting for only 5% of the reported incidents. The youngest age group, 12-17 years, reports the lowest number of crimes at 2.32%.

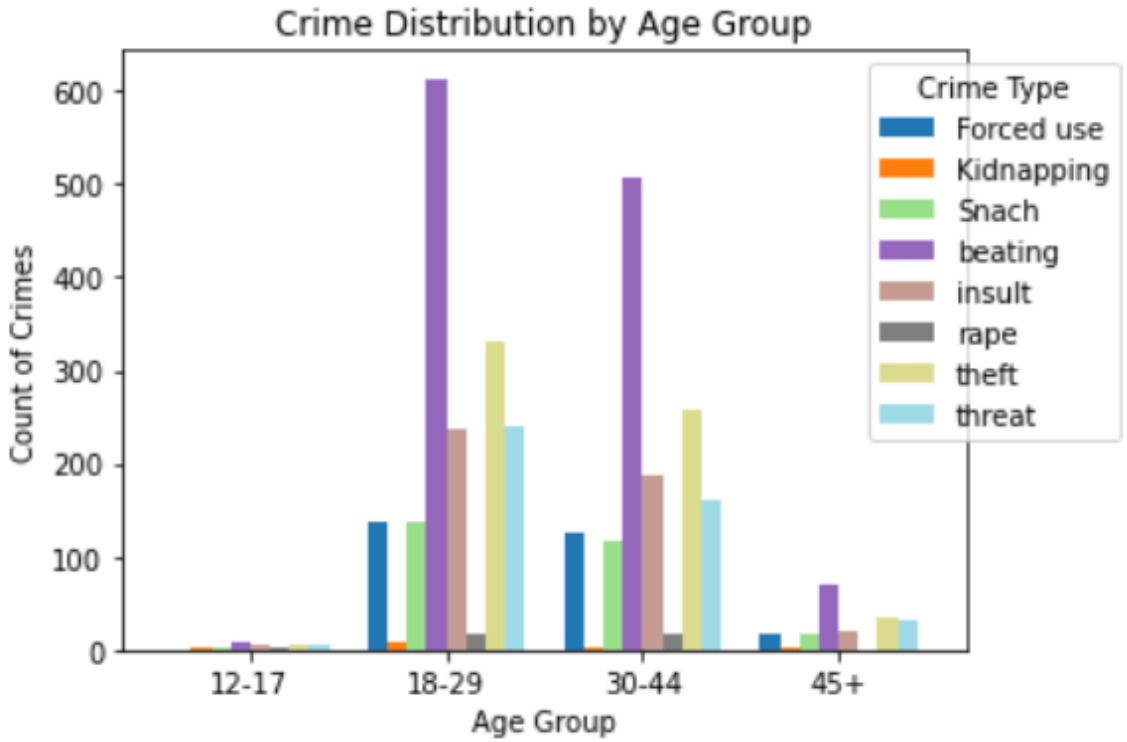

Figure 4. 3 Frequency of crime based on age group

This distribution underscores a concentration of crime incidents among younger adults, particularly those in their late teens to early thirties. It suggests potential trends in criminal activity that may inform targeted interventions or preventive measures aimed at specific age demographics. Further analysis could explore socio-economic factors or environmental influences contributing to these age-specific patterns in crime.

Table 4. 4 Frequency of crime based on demographic area

Subcity or Kebele	Number of Crimes	Percentage of Total (%)
Tabor	851	25.65
Hayk Dar	525	15.82
Bahel Adarash	482	14.53
Menehariya	438	13.2
Adiss Ketema	434	13.08
Mehale	295	8.89
Misrak	293	8.83

The Table 4.4 presents an overview of crime occurrences across various subcities or kebeles based on the analyzed dataset. It shows that Tabor has the highest number of reported crimes, accounting for 25.65% of the total incidents. Following Tabor, Hayk Dar and Bahel Adarash report 15.82% and 14.53% of the crimes, respectively. Menehariya and Adiss Ketema also show notable figures, each contributing around 13% of the total reported crimes. Mehale and Misrak have relatively lower percentages, with 8.89% and 8.83% of the crimes, respectively.

These findings highlight the distribution of crime incidents across different neighborhoods or districts within the study area. Tabor reported crimes, suggesting a potential need for targeted law enforcement or community safety initiatives in that area. Understanding such geographical patterns can assist local authorities in allocating resources more effectively to address specific crime challenges in each subcity or kebele. Further analysis could explore underlying factors contributing to these variations in crime rates among different neighborhoods.

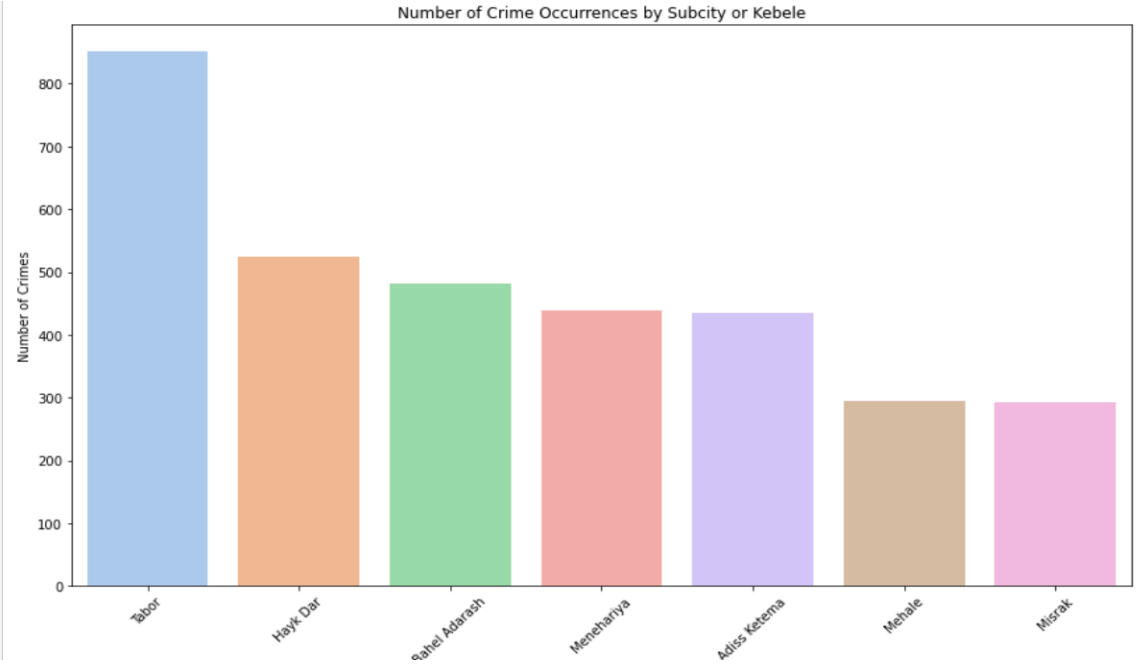

Figure 4. 4 Number of crime occurrence based on demographic area graphically

On the previous section the experiment shows that Random forest got good result so now we will see detain on this algorithm.

On the trained model table result show that 70/30 has better result so the good result selected and with the changing of parameter there is result difference so in n_estimator 50 and 100 accuracy result is 92.87%.

Results Summary:

n_estimators	accuracy
50.0	0.928714859437751
100.0	0.928714859437751
200.0	0.927710843373494
500.0	0.9267068273092369

Table 4. 5 Model Performance Evaluation

Hyperparameter	Parameter	Value	Splitting
N estimator n_estimators = 100	Accuracy	92.87%	70/30
	Ensemble method	Bagging	
	Algorithm type	RandomForest Classifier	
	Number of valid data	3318	
N estimator n_estimators = 100	Accuracy	92%	80/20
	Ensemble method	Bagging	
	Algorithm type	RandomForest Classifier	
	Number of valid data	3318	

Random Forest Classifier:

	precision	recall	f1-score	support
Forced use	1.00	1.00	1.00	53
Kidnapping	0.50	0.50	0.50	2
Snach	0.59	0.39	0.47	61
beating	1.00	1.00	1.00	242
insult	1.00	1.00	1.00	93
rape	0.90	0.90	0.90	10
theft	0.72	0.85	0.78	114
threat	1.00	1.00	1.00	89
accuracy			0.92	664
macro avg	0.84	0.83	0.83	664
weighted avg	0.91	0.92	0.91	664

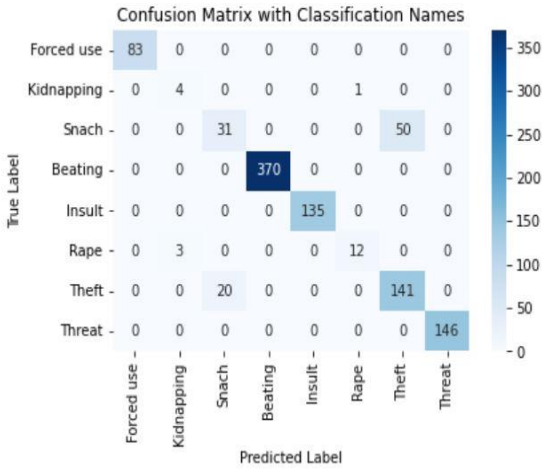

Figure 4. 5 Result of 80/20 classification report and confusion matrix

For 70/30 splitting method the result in on the below

```
# Make predictions on the test set
y_pred = rf_classifier.predict(X_test_normalized)
```

```
# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.928714859437751
```

For 80/20 splitting method the result in on the below

```
# Make predictions on the test set
y_pred = rf_classifier.predict(X_test_normalized)
```

```
# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.9156626506024096
```

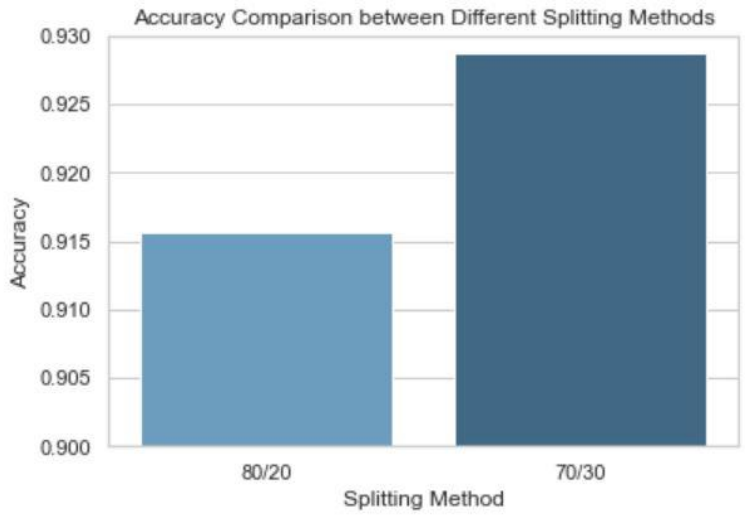

Figure 4. 6 Result for splitting

4.2 Classification Result Report

1. Precision is the ratio of true positives to the sum of true positives and false positives. It measures the accuracy of positive predictions.

2. Recall, alternatively termed sensitivity or true positive rate, represents the proportion of true positives relative to the sum of true positives and false negatives. It measures the model's ability to capture all positive instances.
3. The F1-score is the mean of precision and recall, providing a balance between the two metrics.
4. Support: Indicates the number of actual instances for each class.
5. Accuracy: The percentage of correctly classified instances over the total number of instances.
6. Macro Avg: The unweighted average of precision, recall, and F1-score across all classes.
7. Weighted Avg: The weighted average of precision, recall, and F1-score, where each class's score is weighted by its support.

```
# Print Classification Report
print("Random Forest Classifier:")
print(classification_report(y_test, rf_predictions))
```

```
Random Forest Classifier:
              precision    recall  f1-score   support

  Forced use       1.00      1.00      1.00        83
  Kidnapping       0.57      0.80      0.67         5
       Snach       0.61      0.38      0.47        81
     beating       1.00      1.00      1.00       370
      insult       1.00      1.00      1.00       135
        rape       0.92      0.80      0.86        15
       theft       0.74      0.88      0.80       161
      threat       1.00      1.00      1.00       146

    accuracy                           0.93       996
   macro avg       0.86      0.86      0.85       996
weighted avg       0.92      0.93      0.92       996
```

Figure 4. 7 Classification Report of 70/30

Based on the evaluation metrics provided for the Random Forest Classifier, here are some conclusions:

1. Overall Performance: The classifier achieved an accuracy of 93%, indicating that it correctly classified 93% of the instances in the dataset.
2. Precision, Recall, and F1-score:

For classes like "Forced use," "beating," "insult," and "threat," the classifier achieved very high precision, recall, and F1-scores (all at or close to 1.00), indicating excellent performance in correctly identifying these classes.

However, for classes like "Kidnapping," "Snatch," "rape," and "theft," the performance is mixed. While precision and recall are relatively high for some classes (e.g., "theft"), they are lower for others (e.g., "Kidnapping" and "Snatch").

3. Support: Support refers to the number of instances of each class in the dataset. Classes with lower support may have more variability in their performance metrics due to the smaller sample size.
4. Macro and Weighted Averages: The macro average calculates the metrics independently for each class and then takes the average, giving each class equal weight. The weighted average calculates the metrics for each class and then takes the average, weighted by the number of true instances for each class. In this case, the weighted average is slightly lower than the macro average, indicating some class imbalance because some sensitive data are not easily available and less in number.
5. Conclusion: Overall, the Random Forest Classifier performs very well, especially for classes with higher support. However, there is room for improvement in classes with lower precision, recall, and F1-scores, such as "Kidnapping," "Snatch," "rape," and "theft." Further investigation into these classes, including potentially collecting more data or adjusting the

classifier parameters, may help improve performance for these specific cases.

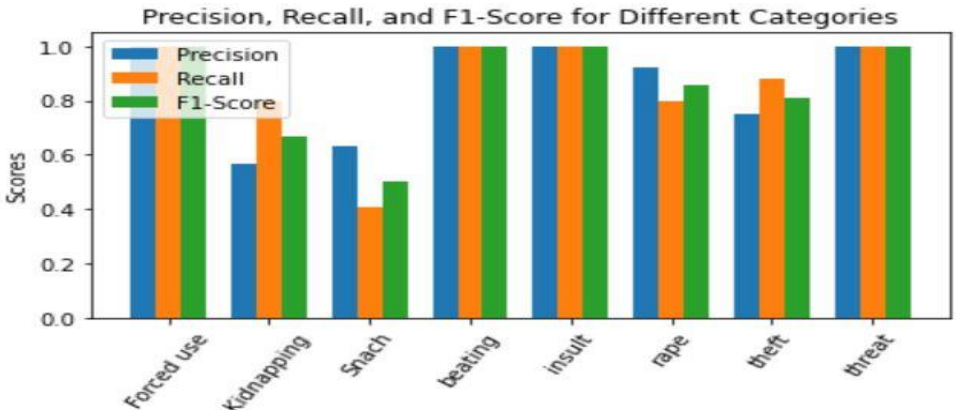

Figure 4. 8 Graphical representation Result Report

Table 4. 6 Final Result Report

Class	Precision(%)	recall(%)	f1-score(%)	support
Forced use	100	100	100	83
Kidnapping	57	80	67	5
Snach	61	38	47	81
Beating	100	100	100	370
Insult	100	100	100	135
Rape	92	80	86	15
Theft	74	88	80	161
Threat	100	100	100	146

Evaluation	Precision(%)	recall(%)	f1-score(%)	Support
Macro avg	86%	86%	85%	996
Weighted avg	92%	93%	92%	996

In summary, this classification report provides a detailed evaluation of the model's performance for each class and overall. It's useful for understanding how well the model is doing across different categories and identifying areas for improvement.

4.3 Confusion Matrix report

Each row represents the actual class, and each column represents the predicted class. Here's a breakdown:

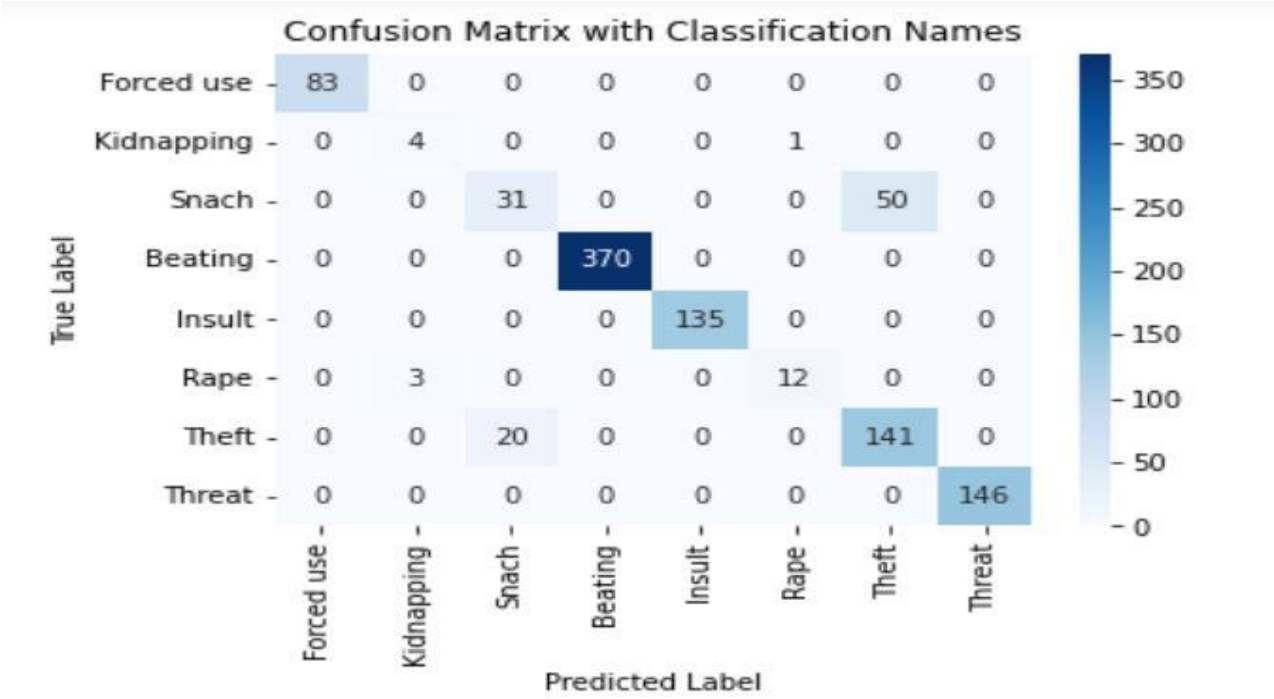

Figure 4. 9 Confusion Matrix Random Forest classifier 70/30

Correctly classified and Misclassified

1. Correctly classified: In classification, the objective is to forecast the categorical class labels of new instances using historical observations.
2. Misclassification: occurs when a model makes an incorrect prediction about the class of an instance. It is the discrepancy between the predicted class and the true class.

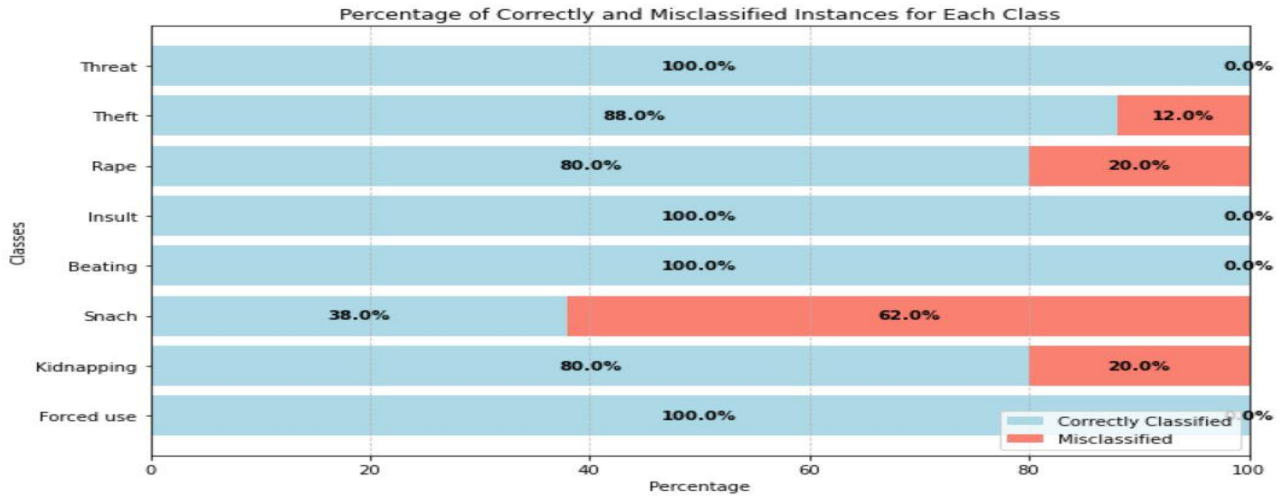

Figure 4. 10 Classified and misclassified

This summary provides Figure 4.10 and Table 4.4 with a clear breakdown for each class, indicating the true positives, total instances, and false positives. It helps assess the model's performance for each category.

Table 4. 7 Clear Breakdown of the Confusion Matrix

Forced use(class 1)	True Positives (TP): 83 Total instances (T): 83 False Positives (F): 0
Kidnapping(class 2)	True Positives (TP): 4 Total instances (T): 5 False Positives (F): 1
Snatch(class 3)	True Positives (TP): 31 Total instances (T): 81 False Positives (F): 50
Beating(class 4)	True Positives (TP): 370 Total instances (T): 370 False Positives (F): 0
Insult(class 5)	True Positives (TP): 135 Total instances(T): 135 False Positives (F): 0

Rape(class 6)	True Positives (TP): 3 Total instances (T): 15 False Positives (F): 12
Theft(class 7)	True Positives (TP): 141 Total instances (T): 161 False Positives (F): 20
Threat(class 8)	True Positives (TP): 146 Total instances (T): 146 False Positives (F): 0

Actaual and Predicted

1. Actual: the actual values are the ground truth or real outcomes of the target variable in your dataset. These are the correct labels that represent the true classes of the instances.
2. Predicted: The predicted values are the outcomes that your machine learning model generates based on its learned patterns from the input features. These values represent the model's predictions for the target variable.

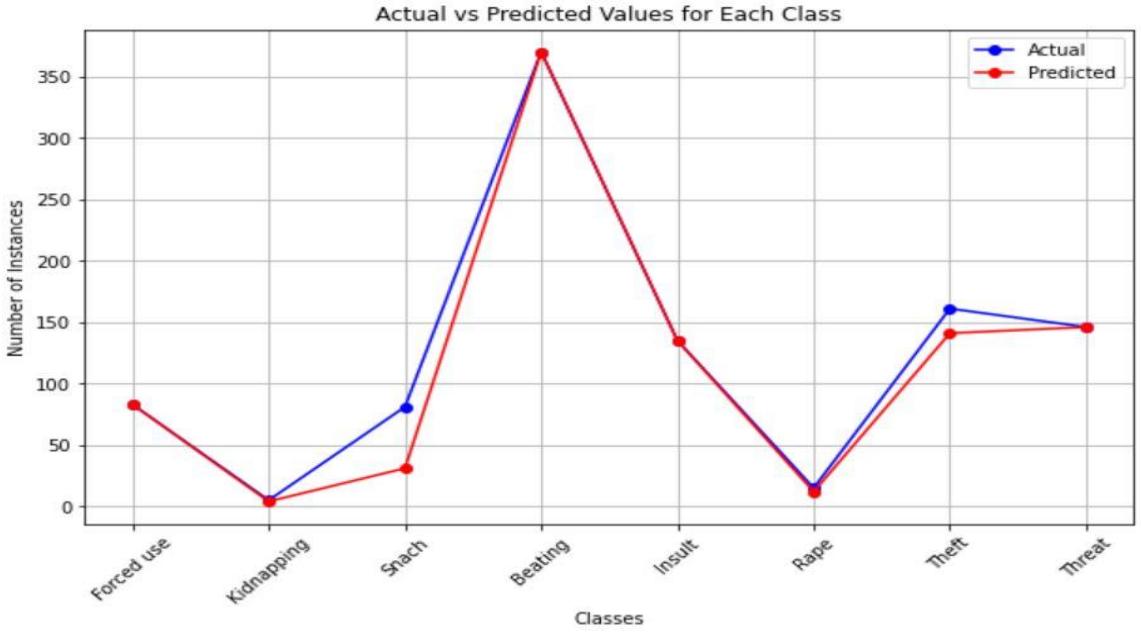

Figure 4. 11 Actual vs. predicted for each class

CHAPTER FIVE

5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

The main objective of this thesis is to propose a model that can predict and analyze crimes against women. To achieve this, an ensemble model with a Random Forest classifier has been successfully developed. The purpose of this model is to serve as a preventive mechanism. It can predict the location and timing of crimes so that women can take necessary precautions to protect themselves before a crime occurs. Based on the data collected, the model has demonstrated strong predictive capabilities, particularly in accurately identifying certain classes with high precision and recall.

Based on the evaluation of machine learning classifiers for predicting crimes against women in Hawassa City, Random Forest emerges as the most effective model with an accuracy of 92.87%, demonstrating strong precision, recall, and F1-scores across multiple crime categories such as "Beating," "Insult," and "Threat." This underscores its capability in accurately classifying prevalent crimes and its suitability for precise crime prediction tasks. Decision Tree follows closely with an accuracy of 91.37%, offering interpretability alongside robust performance metrics. Logistic Regression achieves an accuracy of 91.57%, maintaining high precision and recall rates, making it a reliable choice for binary classification in crime prediction. Support Vector Machine (SVM) also performs consistently well with an accuracy of 90.86%, effectively handling diverse data relationships. The Voting classifier, utilizing ensemble methods, achieves competitive results with an accuracy of 90.86%, demonstrating the effectiveness of model aggregation in enhancing predictive performance. AdaBoost, while slightly lower in accuracy at 89.86%, reveals challenges in achieving high precision particularly for less frequent crime types. These results highlight the diverse strengths of each algorithm and underscore the importance of selecting the appropriate model based on specific predictive needs to improve crime prediction systems and ensure effective safety measures for women in Hawassa City.

- The confusion matrix analysis reveals that, for classes like "Forced use" and "Beating," the model performs exceptionally well with few misclassifications.

- For classes such as "Kidnapping" and "Snatch," there is room for improvement, as reflected in lower precision, recall, and F1-score values.
- The classification report provides detailed metrics for each class, including precision, recall, and F1-score. Notable observations include high precision and recall for classes like "Forced use," "Beating," and "Insult," indicating robust performance in correctly identifying instances of these classes. RandomForestClassifier with 100 hyperparameters got a high accuracy of 92.87%.

5.2 Future work

Based on different data sizes the obtained results of accuracy perform outperforms Random Forest.

This thesis focuses on the city of Hawassa to ensure a more comprehensive study and better results.

- To improve efficiency and effectiveness, it is recommended to prioritize data collection and organization. Implementing automated data collection methods at the Hawassa police station and Women and Children Affairs office would be beneficial, along with working towards a national standard database format for recording crime-related data, and ensuring sustainable maintenance facilities are in place to provide adequate and timely decision-making support.
- Further investigation into misclassified instances, especially for classes with lower precision and recall, may provide insights into potential model improvements and get more data specifically on sensitive crime types.
- Consideration of different hyperparameters, or exploring more algorithms could enhance the model's performance.
- For future making IOT devices (Internet Of Things)interconnected to physical devices, devices used to collect and exchange data, and software for the better safety of women.

REFERENCES

[1] N. Rico, GENDER-BASED VIOLENCE: A HUMAN RIGHTS ISSUE, 1997.

[2] THE CRIMINAL CODE OF THE FEDERAL DEMOCRATIC REPUBLIC OF ETHIOPIA, Ethiopia, Proclamation No.414/2004.

[3] dhsprogram.

[4] U. -HABITAT, The Global Assessment on Women's Safety 1-82.

[5] J. Yin, "Crime Prediction Methods Based on Machine Learning:," October 2022.

[6] Mariette Awad & Rahul Khanna, Efficient Learning Machines, April 2015.

[7] Karabo Jenga, Cagatay Catal & Gorkem Kar , "Machine learning in crime prediction," *Ambient Intelligence and Humanized Computing,* 02 February 2023.

[8] V. Sushma Swaraj, L. Bhavya, G. Pooja, R. DevaRevathi, "Women Safety Prediction using Logistic," *International Journal of Recent Technology and Engineering,* pp. 1-5, March 2020.

[9] Aliasgar Eranpurwala1, Fatema Indorewala2, Nafisa Mapari3, Saloni Mishra, "Women Safety Application for Safe Route Prediction," *International Research Journal of Engineering and Technology (IRJET) ,* pp. 1-5, May 2021.

[10] V. Sushma Swaraj, L. Bhavya, G. Pooja, R. DevaRevathi, "Women Safety Prediction using Logistic," *International Journal of Recent Technology and Engineering (IJRTE),* pp. 1-5, March 2020.

[11] Prof. Shivaprasad More1, Sakshi Mench2,Saloni Kuge3,Hafsa Bagwan4, "Crime Prediction Using Machine Learning," *International Journal of Advanced Research in Computer and Communication Engineering,* pp. 1-5, May 2021.

[12] Varun Mandalapu,Lavanya Elluri,Piyush Vyas and Nirmalya Roy, "Crime Prediction Using Machine Learning," 28 March 2023.

[13] G. Bonaccorso, Machine Learning Algorithms, packt publishing 2017, July 2017.

[14] A. Kumar, Most Common Machine Learning Tasks, https://vitalflux.com/7-common-machine-learning-tasks-related-methods/, Decemeber 4,2022.

[15] Safae Sossi,Brahim Aksasse,Yousef Farhaoui, Data mining and Machine learning Approaches and Technologies For Diagonsing Diabetes in Women, ResearchGates, January 2020.

[16] Manika Lamba,Madhusudhan Margam, "Predictive Modeling," in *Text Mining for Information* , Springer Nature, April 2022.

[17] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research," 22 March 2021.

[18] Yudish Teshal Badal, · Roopesh Kevin Sungkur, "Predictive modelling and analytics of students' grades," 8 September 2022.

[19] Enes Makalic & Daniel Francis Schmidt , "Review of Modern Logistic Regression Methods with Application to Small and Medium Sample Size Problems," 2010.

[20] Daniel Jurafsky & James H. Martin, "Logistic Regression," in *Speech and Language Processing*, Stanford unicersity.https://web.stanford.edu/~jurafsky/slp3/5.pdf, Draft of January 7, 2023..

[21] Emmanuel Ahishakiye,Elisha Opiyo Omulo,Danison Taremwa,Ivan Niyonzima, "Crime Prediction Using Decision Tree (J48)," 03, May 2017.

[22] H. Tan, "Machine Learning Algorithm for Classification," 2021.

[23] R. E. Schapire, Random Forests, 2001 Kluwer Academic Publishers. Manufactured in The Netherlands springer.com.

[24] Saso Dzeroski, Pance Panov & Bernard Zenko, "Machine Learning, Ensemble Methods," in *Encyclopedia of Complexity and Systems Science*, https://link.springer.com/referenceworkentry/10.1007/978-0-387-30440-3_315, 2009.

[25] L. Rokach, "ENSEMBLE METHODS FOR CLASSIFIERS," in *Data Mining and Knowledge Discovery Handbook*, 2005.

[26] Assistant Professor, School of Legal Studies, HPU Regional Centre, Khanyara, Dharamshal, H.P, "Criminology: A Study on Crime against Women," vol. Index Copernicus Value (2016), 2017.

[27] V. Sushma Swaraj,L. Bhavya,G. Pooja,R. Deva Revathi, "Women Safety Prediction using Logistic Regression Model," *International Journal of Recent Technology and Engineering (IJRTE) ResearchGate,* Vols. Volume-8, no. Issue-6, p. , March 2020.

[28] Tanya SIngh,Namya H,Suchanda Dutta,Tirumala Reddy,Dr.Manimozhi I,Prof. Neha Gopal N, "ANALYSIS, FORECASTING AND PREDICTION OF CRIME AGAINST WOMEN USING MACHINE

LEARNING TECHNIQUES," *International jornal of novel research and developement,* Vols. Volume 8,, no. Issue 5, May 2023.

[29] B. Z. Wubineh, "Crime analysis and prediction using machine-learning approach in the case of Hossana Police Commission," *Security Journa,Springer Nature,* March 2024.

[30] Vinay Narayan Bhat,V Santhosh Kumar,Prof. Saravanan C,, "Analysis and Prediction of Crime against women in India using machine learning algorithms.," *Journal of Emerging Technologies and Innovative Research,* vol. Volume 8, no. Issue 6, June 2021.

[31] Purvi Prasad,Amrita Nair,Dr. S. Godfrey Winster, "Crime Against Women: Analysis And Prediction," *International Journal of Engineering Research & Technology,* vol. Vol. 10, no. Issue 05,, May-2021.

[32] S. Lavanyaa, D. Akila, "Crime against Women (CAW) Analysis and Prediction in Tamilnadu Police Using Data Mining Techniques," *International jornal of recent technology and engineering ,* vol. volume 7, no. issue 5, February 2019.

[33] W. S. V. Lakshan,A. T. P. Silva,W. A. C. Weerakoon, "An enhanced ensemble model for crime occurrence prediction," 2021.

[34] S. M. S. Kabir, Basic Guidelines for Research: An Introductory Approach for All Disciplines, July 2016.

[35] A. Brijith, "Data Preprocessing for Machine Learning," no. https://www.researchgate.net/publication/375003512, pp. 1-5, · October 2023.

[36] Kiran Maharana,Surajit Mondal,Bhushankumar Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings,* pp. 1-9, 2022.

[37] A. Sojasingarayar, "https://medium.com/," 3 nov 2022. [Online]. Available: https://medium.com/@abonia/data-preprocessing-in-python-1f90d95d44f4.

[38] "Key Machine Learning Concepts Explained — Dataset Splitting and Random Forest," 1 FEBRUARY 2020. [Online]. Available: https://www.freecodecamp.org/news/key-machine-learning-concepts-explained-dataset-splitting-and-random-forest/.

[39] H. Dalianis, "Evaluation Metrics and Evaluation," in *Clinical Text Mining*, springer, 15 May 2018.

[40] Mirka Saarela , Susanne Jauhiainen, "Comparison of feature importance measures as explanations," no. SN Applied Sciences (2021) 3:272 | https://doi.org/10.1007/s42452-021-04148-9, pp. 1-12, 3 February 2021.

[41] "Geeks for Geeks," [Online]. Available: https://www.geeksforgeeks.org/feature-importance-with-random-forests/.

[42] Hamzah A. Alsayadi,Nima Khodadadi,Sunil Kumar, "Improving the Regression of Communities and Crime Using Ensemble of Machine Learning Models," Vols. Vol. 01, No. 01, PP. 27-34, August 2022.

[43] W. S. V. Lakshan,A. T. P. Silva,W. A. C. Weerakoon, "An enhanced ensemble model for crime occurrence prediction," Sri Lanka , 2021.

[44] R. Karthik Sriraam,S.M. Keerthivasan, K. Sukant,A. Krishnamoorthy, "Crime Prediction and Analysis," 2022.

[45] Yasmine Lamari , Bartol Freskura , Anass Abdessamad , Sarah Eichberg and Simon de Bonville, "Predicting Spatial Crime Occurrences through an," 29 October 2020.

[46] Women Safety Prediction using Logistic Regression Model V. Sushma Swaraj, L. Bhavya, G. Pooja, R. DevaRevathi, "Women Safety Prediction using Logistic," Vols. Volume-8 Issue-6,, March 2020.

[47] Vinay Narayan Bhat,V Santhosh Kumar,Prof. Saravanan C, "Analysis and Prediction of Crime against women in," Vols. Volume 8,, no. Issue 6 , June 2021.

[48] V. Sushma Swaraj, L. Bhavya, G. Pooja, R. DevaRevathi, "Women Safety Prediction using Logistic," Vols. Volume-8 , no. Issue-6, March 2020.

[49] Aliasgar Eranpurwala, Fatema Indorewala, Nafisa Mapari, Saloni Mishra, "Women Safety Application for Safe Route Prediction," vol. Volume: 08 , no. Issue: 05 , May 2021.

[50] S. Lavanyaa, D. Akila, "Crime against Women (CAW) Analysis and," Vols. Volume-7, no. Issue-5C, February 2019.

[51] Purvi Prasad,Amrita Nair,Dr. S. Godfrey Winster, "Crime Against Women: Analysis And Prediction," vol. Vol. 10 , no. Issue 05, May-2021.

[52] Ismail Olaniyi Muraina, "IDEAL DATASET SPLITTING RATIOS IN MACHINE LEARNING ALGORITHMS:," in *https://www.researchgate.net/publication/358284895*, February 2022.

[53] Praveen Boinee, Alessandro De Angelis, and Gian Luca Foresti , "Meta Random Forests," January 2006.

APPENDIX

A-Data loading and selection of algorithm

```
# Load your data
df = pd.read_csv("veronica.csv")

# Extract features and target variable
X = df[['Subcity', 'kebele', 'Year', 'Age', 'Maritalstatus', 'Timeofcrimehappen ', 'categories']]
y = df['Crime ']

# Convert categorical variables to one-hot encoding
X = pd.get_dummies(X, columns=['Subcity', 'kebele', 'Year', 'Age', 'Maritalstatus', 'Timeofcrimehappen ', 'categories'])

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Normalize the features using StandardScaler
scaler = StandardScaler()
X_train_normalized = scaler.fit_transform(X_train)
X_test_normalized = scaler.transform(X_test)

# Initialize classifiers of interest
classifiers = {
    'Decision Tree': DecisionTreeClassifier(random_state=42),
    'Voting': VotingClassifier(estimators=[
        ('dt', DecisionTreeClassifier(random_state=42)),
        ('lr', LogisticRegression(random_state=42, max_iter=1000)),
        ('svc', SVC(random_state=42, probability=True)),
        ('gnb', GaussianNB())
    ], voting='soft'),
    'Random Forest': RandomForestClassifier(random_state=42, n_estimators=100),
    'AdaBoost': AdaBoostClassifier(base_estimator=DecisionTreeClassifier(random_state=42, max_depth=1), n_estimators=50, random_s
    'SVM': SVC(random_state=42),
    'Logistic Regression': LogisticRegression(random_state=42, max_iter=1000)
}
```

B-During Training of algorithm

```
# Convert categorical variables to one-hot encoding
X = pd.get_dummies(X, columns=['Subcity', 'kebele', 'Year', 'Age', 'Maritalstatus', 'Timeofcrimehappen ', 'categories'])

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Normalize the features using StandardScaler
scaler = StandardScaler()
X_train_normalized = scaler.fit_transform(X_train)
X_test_normalized = scaler.transform(X_test)

# Initialize classifiers of interest
classifiers = {
    'Decision Tree': DecisionTreeClassifier(random_state=42),
    'Voting': VotingClassifier(estimators=[
        ('dt', DecisionTreeClassifier(random_state=42)),
        ('lr', LogisticRegression(random_state=42, max_iter=1000)),
        ('svc', SVC(random_state=42, probability=True)),
        ('gnb', GaussianNB())
    ], voting='soft'),
    'Random Forest': RandomForestClassifier(random_state=42, n_estimators=100),
    'AdaBoost': AdaBoostClassifier(base_estimator=DecisionTreeClassifier(random_state=42, max_depth=1), n_estimators=50, random_s
    'SVM': SVC(random_state=42),
    'Logistic Regression': LogisticRegression(random_state=42, max_iter=1000)
}

# Initialize empty lists to store metrics for each classifier
metrics = []
```

C-Performance metrics of algorithm

```
# Evaluate each classifier of interest
for clf_name, clf in classifiers.items():
    # Train the classifier
    clf.fit(X_train_normalized, y_train)

    # Predictions
    y_pred = clf.predict(X_test_normalized)

    # Calculate metrics
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred, average='weighted')
    recall = recall_score(y_test, y_pred, average='weighted')
    f1 = f1_score(y_test, y_pred, average='weighted')

    # Append metrics to list, converting to whole numbers
    metrics.append({
        'Classifier': clf_name,
        'Accuracy (%)': accuracy * 100,  # Convert to percentage
        'Precision (%)': precision * 100,  # Convert to percentage
        'Recall (%)': recall * 100,  # Convert to percentage
        'F1-score (%)': f1 * 100  # Convert to percentage
    })

# Convert metrics list to a pandas DataFrame
metrics_df = pd.DataFrame(metrics)

# Display the metrics table with lines between rows
print("Performance Metrics of Classification Algorithms:")
print(tabulate(metrics_df, headers='keys', tablefmt='fancy_grid', showindex=False, numalign='center'))
```