Crime Data Analysis

*A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree*

*of*

**Bachelor of Technology**

**in The Department of CSE**

**Big Data Analytics- 22DSB3303A**

Submitted by

**2210030010: Meghana L**

**2210030330: Sohana Reddy P**

**2210030060: Bhavana S**

Under the guidance of

**Dr. Shahin Fatima**



Department of Electronics and Communication Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075 (Optional)

MAY - 2025.

**Abstract**

Crime is a growing societal issue affecting public safety, resource allocation, and city development. With the rise in available data and technological tools, data analytics offers new opportunities to understand crime patterns, detect anomalies, and implement proactive measures. This project explores various machine learning techniques for analyzing crime datasets and predicting crime types and locations using clustering, classification, and anomaly detection methods.

Crime data analysis plays an increasingly vital role in law enforcement, urban planning, and policymaking by offering a scientific approach to understanding and predicting criminal behavior. With the proliferation of open-source crime datasets and advancements in machine learning, authorities can now uncover patterns that were previously hidden in massive data repositories. This transformation in crime analytics allows law enforcement to not only react to incidents but also anticipate and prevent potential threats.

**Crime Data Analysis**

**Introduction**

Crime data analysis plays an increasingly vital role in law enforcement, urban planning, and policymaking by offering a scientific approach to understanding and predicting criminal behavior. With the proliferation of open-source crime datasets and advancements in machine learning, authorities can now uncover patterns that were previously hidden in massive data repositories. This transformation in crime analytics allows law enforcement to not only react to incidents but also anticipate and prevent potential threats.

Modern crime data analysis employs a wide range of techniques, from simple visualizations to complex machine learning models. Clustering techniques like K-Means are used to identify crime hotspots, while classification algorithms such as Naïve Bayes help in categorizing crime types. Anomaly detection techniques, such as Z-Score and Interquartile Range (IQR), are utilized to detect outliers that could signal unusual criminal activity.

This project delves into these methodologies, emphasizing their application in real-world scenarios using publicly available datasets, such as those from Kaggle. The insights derived from such analysis not only support informed decision-making but also improve the strategic deployment of security forces and resources. Ultimately, this empowers policymakers and enforcement agencies with actionable intelligence to enhance public safety.

**PROBLEM STATEMENT**

The traditional approach to handling crime data is often reactive, relying on manual analysis and reporting. This method struggles to keep pace with the massive volume of data generated daily across regions. Additionally, unstructured data formats and the lack of automated pattern recognition hinder efficient resource planning and crime prevention.

The objective of this project is to implement a crime data analysis system that utilizes data analytics techniques such as clustering, classification, and anomaly detection to uncover patterns, forecast criminal activities, and support better decision-making. This system aims to:

* Detect crime-prone areas using clustering methods.
* Categorize types of crimes using classification techniques.
* Identify outlier incidents through anomaly detection.

**METHODOLOGY**

**Dataset**

The dataset used in this project was sourced from Kaggle and contains information such as crime types, locations, time, and dates. The data was cleaned and pre-processed to ensure consistency and accuracy.

1. **Techniques Used**
2. **K-Means Clustering** – to identify and visualize crime hotspots.
3. **Z-Score and IQR** – to perform anomaly detection for identifying unusual events.
4. **Naïve Bayes Classifier** – to categorize crimes based on different features.
5. **Tools & Technologies**

* R
* R Studio
* Dplyr, ggplot2, caret, lubridate

The following flow was followed:

1. Data Preprocessing
2. Feature Selection
3. Model Implementation
4. Visualization of Results

A screenshot of a computer

AI-generated content may be incorrect.

**METHODOLOGY**

**Code**

library(dplyr)

library(lubridate)

library(e1071)

library(caret)

library(ggplot2)

crime\_data <- read.csv("C:/Users/Asus/Downloads/BDA/crime\_dataset\_india.csv", stringsAsFactors = FALSE)

crime\_data$Date\_Reported <- mdy\_hm(crime\_data$Date.Reported)

crime\_data$Date\_of\_Occurrence <- mdy\_hm(crime\_data$Date.of.Occurrence)

crime\_data$Date\_Case\_Closed <- mdy\_hm(crime\_data$Date.Case.Closed)

crime\_data$Month <- month(crime\_data$Date\_of\_Occurrence)

crime\_data$Weekday <- weekdays(crime\_data$Date\_of\_Occurrence)

crime\_data$Case\_Closed\_Flag <- ifelse(tolower(crime\_data$Case.Closed) == "yes", 1, 0)

model\_data <- crime\_data %>%

select(City, Crime.Code, Victim.Age, Victim.Gender, Weapon.Used, Police.Deployed, Case\_Closed\_Flag) %>%

na.omit()

model\_data$City <- as.factor(model\_data$City)

model\_data$Crime.Code <- as.factor(model\_data$Crime.Code)

model\_data$Victim.Gender <- as.factor(model\_data$Victim.Gender)

model\_data$Weapon.Used <- as.factor(model\_data$Weapon.Used)

model\_data$Case\_Closed\_Flag <- as.factor(model\_data$Case\_Closed\_Flag)

set.seed(123)

trainIndex <- createDataPartition(model\_data$Case\_Closed\_Flag, p = 0.8, list = FALSE)

train\_data <- model\_data[trainIndex, ]

test\_data <- model\_data[-trainIndex, ]

nb\_model <- naiveBayes(Case\_Closed\_Flag ~ ., data = train\_data)

predictions <- predict(nb\_model, test\_data)

confusionMatrix(predictions, test\_data$Case\_Closed\_Flag)

z\_scores <- scale(crime\_data$Victim.Age)

crime\_data$Z\_Anomaly\_Age <- ifelse(abs(z\_scores) > 3, 1, 0)

anomalies\_age <- crime\_data[crime\_data$Z\_Anomaly\_Age == 1, ]

Q1 <- quantile(crime\_data$Police.Deployed, 0.25, na.rm = TRUE)

Q3 <- quantile(crime\_data$Police.Deployed, 0.75, na.rm = TRUE)

IQR\_val <- Q3 - Q1

lower\_bound <- Q1 - 1.5 \* IQR\_val

upper\_bound <- Q3 + 1.5 \* IQR\_val

crime\_data$IQR\_Anomaly\_PD <- ifelse(crime\_data$Police.Deployed < lower\_bound | crime\_data$Police.Deployed > upper\_bound, 1, 0)

anomalies\_pd <- crime\_data[crime\_data$IQR\_Anomaly\_PD == 1, ]

unsolved\_by\_city <- crime\_data %>%

filter(Case\_Closed\_Flag == 0) %>%

count(City, sort = TRUE)

crime\_data$Hour <- substr(crime\_data$Time.of.Occurrence, 12, 13)

table(crime\_data$Hour)

crime\_data %>%

mutate(Year = year(Date\_Reported)) %>%

group\_by(Year) %>%

summarise(Total\_Crimes = n()) -> crime\_summary

ggplot(crime\_summary, aes(x = Year, y = Total\_Crimes)) +

geom\_line(color = "blue", size = 1.2) +

geom\_point(color = "red", size = 3) +

labs(title = "Crime Rate Comparison by Year",

x = "Year",

y = "Total Crimes Reported") +

scale\_y\_continuous(breaks = seq(0, 5000, by = 500), limits = c(0, 5000)) +

theme\_minimal()

unresolved\_cases <- crime\_data %>%

filter(tolower(Case.Closed) == "no") %>%

group\_by(City) %>%

summarise(Unresolved\_Count = n()) %>%

mutate(Percentage = round(100 \* Unresolved\_Count / sum(Unresolved\_Count), 1),

Label = paste0(City, " (", Percentage, "%)"))

ggplot(unresolved\_cases, aes(x = "", y = Percentage, fill = Label)) +

geom\_bar(stat = "identity", width = 1) +

coord\_polar(theta = "y") +

labs(title = "Percentage of Unresolved Cases by City") +

theme\_void() +

theme(legend.title = element\_blank())

top\_crimes\_by\_city <- crime\_data %>%

group\_by(City, Crime.Description) %>%

summarise(Frequency = n(), .groups = 'drop\_last') %>%

slice\_max(order\_by = Frequency, n = 1)

ggplot(top\_crimes\_by\_city, aes(x = reorder(City, -Frequency), y = Frequency, fill = Crime.Description)) +

geom\_bar(stat = "identity") +

labs(title = "Most Frequent Crime in Each City",

x = "City",

y = "Number of Crimes",

fill = "Crime Type") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

# **RESULTS**

The application of data analytics methods produced interpretable and valuable results. The K-Means clustering model effectively revealed crime hotspots, especially in urban zones. The anomaly detection method identified critical deviations in patterns that could indicate potential crime spikes. Lastly, the Naïve Bayes classifier successfully categorized crime types with a reasonable level of accuracy.

These outcomes affirm the effectiveness of machine learning in crime analysis. The models can be further tuned with more granular and real-time data for improved performance.

A graph with a line and numbers

AI-generated content may be incorrect.

A colorful circle with numbers and text

AI-generated content may be incorrect.

A graph of crime type

AI-generated content may be incorrect.

# **CONCLUSION and FUTURE WORK**

This project demonstrated the power of integrating data analytics into crime analysis. By leveraging clustering, classification, and anomaly detection techniques, we successfully analyzed crime data to derive actionable insights. These models can support law enforcement agencies in making proactive decisions and allocating resources efficiently.

In the future, this work can be extended by:

* Incorporating real-time crime feeds for dynamic analysis.
* Using advanced models like Random Forest or Neural Networks for higher accuracy.
* Developing a web-based dashboard for interactive crime data visualization.

##### **References**

1. G Kaggle Crime Dataset – https://www.kaggle.com/datasets
2. McKinney, W. (2010). *Data Structures for Statistical Computing in Python*.
3. Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research.
4. Jain, A.K. (2010). *Data Clustering: 50 Years Beyond K-Means*.