**Phase 4.4**

**CAPSTONE PROJECT**

**Crime Data Analysis Project**

**Table of Contents**

**Introduction**

**1.Phase 1**

1.1 Data Collection and Preparation

1.2 Data Sources

**2.Phase 2**

2.1 Analysis of Literacy Rate vs Total Crimes

2.2 Analysis of the Type of Crime vs Each State vs Literacy Rate

2.3 Analysis of Year-on-Year Total Crime Rate

2.4 Analysis of Area vs Overall Crime

2.5 Analysis of Population vs Overall Crime

2.6 State Crime Reports

**3.Phase 3**

3.1 Inserting Records into SQL Tables

3.2 SQL Queries and Analysis

**4.Phase 4**

4.1 Creating Clusters

4.2 DataFrame for Each Cluster

4.3 Analysis and Report

**5.Conclusion**

**6.References**

**Introduction**

This project focuses on the analysis of crime data across various states and union territories (UTs) in India. The goal is to gain insights into crime patterns, understand the impact of literacy rates, population, and area on crime rates, and identify the most and least crime-affected districts. The project also includes SQL operations for data manipulation and retrieval, and the application of unsupervised machine learning techniques to classify districts into sensitive, moderate, and peaceful areas.

**1. PHASE 1 Data Collection**

**1.1 Data Sources**

Data was collected manually from various online sources, including government websites and statistical reports. The primary datasets used in this project are:

* Crime data by district and year
* Literacy rates by state
* Population data by state
* Area data by state

**1.2 Data Collection and Preparation**

The collected data was cleaned to remove any inconsistencies and ensure accuracy. Steps included:

* Removing rows with missing or incomplete data.
* Aggregating data to the state level for state/UT wise analysis.
* Doing Interpolation to fill the missing data

**2. PHASE 2 Analysis**

**2.1 Analysis of Literacy Rate vs Total Crimes**

This analysis explores the relationship between literacy rates and total crimes in each state. By plotting literacy rates against total crimes, we can identify any correlation between these variables.

**2.2 Analysis of the Type of Crime vs Each State vs Literacy Rate**

This section examines how different types of crimes are distributed across states and how literacy rates might influence these distributions. Bar plots and heatmaps are used to visualize this data.

**2.3 Analysis of Year-on-Year Total Crime Rate**

Year-on-year analysis helps in understanding trends in crime rates over time. This involves plotting the total crime rates for each year and observing the changes.

**2.4 Analysis of Area vs Overall Crime**

Here, we analyze how the geographic area of a state influences the overall crime rate. Larger areas may have different crime dynamics compared to smaller areas.

**2.5 Analysis of Population vs Overall Crime**

This section explores the relationship between state population and overall crime rate. Population density can significantly impact crime rates, and this analysis helps in understanding this relationship.

**2.6 State Crime Reports**

Detailed reports are generated for each state, summarizing the findings from the above analyses. These reports include graphical representations and interpretations of the data.

**3. PHASE 3 SQL Operations**

**3.1 Inserting Records into SQL Tables**

Records from the provided crime datasets were inserted into SQL tables. Separate tables were created for each dataset to facilitate efficient data management and retrieval.

**3.2 SQL Queries and Analysis**

Several SQL queries were executed to extract meaningful insights:

* Identifying the highest and lowest number of rapes and kidnappings by state, district, and year.
* Finding the districts with the highest dacoity/robbery incidents.
* Identifying districts with the lowest number of murders.
* Sorting the number of murders in ascending order by district and year.

The results of these queries were used for further analysis and reporting.

**4. PHASE 4 Unsupervised Machine Learning (Clustering)**

**4.1 Creating Clusters**

Using the crime data, districts were clustered into three categories:

* Sensitive Areas
* Moderate Areas
* Peaceful Areas

K-means clustering was employed to classify the districts based on their crime rates.

**4.2 DataFrame for Each Cluster**

Separate DataFrames were created for each cluster, containing data specific to the sensitive, moderate, and peaceful areas.

**4.3 Analysis and Report**

A detailed analysis was conducted on each cluster, focusing on:

* Factors contributing to high crime rates in sensitive areas.
* Measures needed to reduce crime rates in sensitive areas.
* Identifying the most and least safe districts.
* Plotting various graphs to visualize and support the analysis.

**Observations:**

**Sensitive Areas:**

* High population density and lower literacy rates contribute significantly to higher crime rates.
* Urbanization and economic disparity also play crucial roles.
* Measures such as improving education, providing employment opportunities, and enhancing law enforcement can help reduce crime.
* We can see the most dangerous area is Mumbai in Maharashtra.

**Moderate Areas:**

* These areas show balanced crime rates, influenced by moderate literacy rates and average economic conditions.
* Strengthening community policing and promoting social welfare programs can maintain and improve safety.

**Peaceful Areas:**

* Higher literacy rates and lower population density correlate with lower crime rates.
* These areas can serve as models for other regions.
* Maintaining high education standards and ensuring equitable economic development are key to preserving peace.
* We can find the most peaceful place in having less crime rate is CID which is in Manipur.

**Graphs and Visualizations:**

Scatter plots, bar charts, and heatmaps were used to visualize the relationships between various factors and crime rates.

**5. Conclusion**

The project provides a comprehensive analysis of crime data across Indian states and UTs. Insights gained from this analysis can help in formulating targeted policies to reduce crime rates and improve public safety. The use of machine learning for clustering districts into sensitive, moderate, and peaceful areas offers a novel approach to crime analysis, highlighting the importance of data-driven decision-making. By leveraging detailed data on literacy rates, population, and area, this project underscores the multifaceted nature of crime and the need for holistic strategies to address it effectively. Implementing educational, economic, and law enforcement improvements based on these insights can significantly enhance societal well-being and safety.

**6. References**

* National Crime Records Bureau (NCRB)
* Census of India
* Wikipedia - List of Indian States and Union Territories by Literacy Rate
* Government of India - Ministry of Home Affairs

This detailed analysis and the methodologies employed provide a robust framework for understanding and addressing crime in India. The project emphasizes the significance of integrating various data sources and analytical techniques to derive actionable insights.