**Clustering Crime Data for Districts in India**

To create meaningful clusters of districts based on crime data, we'll follow a structured approach using unsupervised machine learning techniques, specifically clustering. We aim to classify districts into three clusters: Sensitive Areas, Moderate Areas, and Peaceful Areas. Here's a step-by-step guide on how to achieve this:

**Step 1: Data Preparation**

**1.1 Collecting Data**

We assume various crime datasets have been provided, including data on crimes against women, crimes against Scheduled Tribes (ST), and IPC crimes. The key columns we will focus on include:

* District
* State
* Year
* Total Crimes
* Crimes against women (e.g., Rape, Kidnapping)
* Crimes against ST (e.g., Dacoity, Robbery)
* IPC crimes (e.g., Murder, Attempt to Murder, Theft)

**1.2 Merging Data**

We will merge these datasets on common columns such as District, State, and Year to create a comprehensive dataset that covers all crime types across all districts.

python

Copy code

import pandas as pd

# Load datasets

women\_crimes = pd.read\_csv('42\_District\_wise\_crimes\_committed\_against\_women\_2001\_2012.csv')

st\_crimes = pd.read\_csv('02\_District\_wise\_crimes\_committed\_against\_ST\_2001\_2012.csv')

ipc\_crimes = pd.read\_csv('01\_District\_wise\_crimes\_committed\_IPC\_2001\_2012.csv')

# Merge datasets

merged\_data = women\_crimes.merge(st\_crimes, on=['State', 'District', 'Year'], suffixes=('\_women', '\_st'))

merged\_data = merged\_data.merge(ipc\_crimes, on=['State', 'District', 'Year'])

**1.3 Feature Engineering**

We'll create new features that might help in clustering, such as the total number of violent crimes, total number of non-violent crimes, and overall crime rate per 100,000 people.

python

Copy code

# Calculate total crimes and crime rates

merged\_data['Total\_Crimes'] = merged\_data['Rape'] + merged\_data['Kidnapping'] + merged\_data['Dacoity'] + merged\_data['Robbery'] + merged\_data['Murder'] + merged\_data['Attempt\_to\_Murder']

merged\_data['Crime\_Rate'] = (merged\_data['Total\_Crimes'] / merged\_data['Population']) \* 100000

**Step 2: Clustering**

**2.1 Normalizing Data**

Since the data might have varying scales, we normalize it before applying clustering algorithms.

python

Copy code

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

scaled\_data = scaler.fit\_transform(merged\_data[['Rape', 'Kidnapping', 'Dacoity', 'Robbery', 'Murder', 'Attempt\_to\_Murder', 'Total\_Crimes', 'Crime\_Rate']])

**2.2 Applying K-Means Clustering**

We'll use K-Means clustering to classify the districts into three clusters.

python

Copy code

from sklearn.cluster import KMeans

# Apply K-Means

kmeans = KMeans(n\_clusters=3, random\_state=42)

merged\_data['Cluster'] = kmeans.fit\_predict(scaled\_data)

**2.3 Labeling Clusters**

After clustering, we label the clusters based on their characteristics (e.g., average crime rate).

python

Copy code

# Assign meaningful names to clusters

cluster\_labels = {0: 'Sensitive Areas', 1: 'Moderate Areas', 2: 'Peaceful Areas'}

merged\_data['Cluster\_Label'] = merged\_data['Cluster'].map(cluster\_labels)

**Step 3: DataFrame for Each Cluster**

**3.1 Creating DataFrames**

We create separate DataFrames for each cluster to facilitate further analysis.

python

Copy code

sensitive\_areas = merged\_data[merged\_data['Cluster\_Label'] == 'Sensitive Areas']

moderate\_areas = merged\_data[merged\_data['Cluster\_Label'] == 'Moderate Areas']

peaceful\_areas = merged\_data[merged\_data['Cluster\_Label'] == 'Peaceful Areas']

**Step 4: Analysis and Reporting**

**4.1 Analyzing Clusters**

We analyze each cluster to identify patterns and factors contributing to crime rates.

**Sensitive Areas:**

* **High Crime Rates**: Districts in this cluster have significantly higher crime rates.
* **Types of Crimes**: High incidents of violent crimes such as murders and rapes.
* **Contributing Factors**: Socio-economic factors, lack of effective law enforcement, and lower literacy rates might be contributing to higher crime rates.

**Moderate Areas:**

* **Moderate Crime Rates**: These districts have average crime rates.
* **Types of Crimes**: A mix of violent and non-violent crimes.
* **Contributing Factors**: Better socio-economic conditions and moderate law enforcement effectiveness.

**Peaceful Areas:**

* **Low Crime Rates**: These districts have significantly lower crime rates.
* **Types of Crimes**: Fewer violent crimes, more incidents of non-violent crimes.
* **Contributing Factors**: Higher literacy rates, effective law enforcement, and better socio-economic conditions.

**4.2 Recommendations**

To reduce crime in sensitive areas:

* **Improving Law Enforcement**: Increase police presence and improve the efficiency of the legal system.
* **Socio-Economic Development**: Focus on education, employment opportunities, and poverty alleviation.
* **Community Programs**: Implement community policing and awareness programs to involve citizens in crime prevention.

**4.3 Identifying Safe and Unsafe Districts**

* **Most Unsafe Districts**: Districts with the highest crime rates and incidents of violent crimes.
* **Most Safe Districts**: Districts with the lowest crime rates and minimal incidents of violent crimes.

**4.4 Visualization**

We use appropriate graphs to visualize the data and support our findings.

python

Copy code

import matplotlib.pyplot as plt

import seaborn as sns

# Crime rates by cluster

plt.figure(figsize=(10, 6))

sns.boxplot(x='Cluster\_Label', y='Crime\_Rate', data=merged\_data)

plt.title('Crime Rates by Cluster')

plt.xlabel('Cluster')

plt.ylabel('Crime Rate per 100,000 people')

plt.show()

# Types of crimes in sensitive areas

sensitive\_crimes = sensitive\_areas[['Rape', 'Kidnapping', 'Dacoity', 'Robbery', 'Murder', 'Attempt\_to\_Murder']].sum()

sensitive\_crimes.plot(kind='bar', figsize=(10, 6), title='Types of Crimes in Sensitive Areas')

plt.xlabel('Type of Crime')

plt.ylabel('Number of Incidents')

plt.show()

**Capstone Project Story (1000 words)**

**Introduction**

Crime is a significant concern for any society, and understanding its patterns and causes is crucial for effective prevention and management. This project aims to classify districts in India into three categories based on their crime data: Sensitive Areas, Moderate Areas, and Peaceful Areas. Using unsupervised machine learning techniques, we provide insights that can help authorities take targeted actions to reduce crime and improve public safety.

**Data Collection and Preparation**

The project started with collecting various crime datasets, including crimes against women, crimes against Scheduled Tribes (ST), and Indian Penal Code (IPC) crimes. These datasets were merged to create a comprehensive dataset that covers all crime types across all districts. Key columns included district, state, year, total crimes, specific crime types (e.g., rape, kidnapping, dacoity, robbery, murder), and population.

**Feature Engineering**

We created new features to help in clustering. Total crimes were calculated as the sum of all crime types, and the crime rate per 100,000 people was computed to standardize the data. These features allowed for a more accurate clustering process.

**Clustering Process**

Normalization was performed on the data to ensure that features with different scales did not distort the clustering results. We then applied K-Means clustering to classify districts into three clusters. The clusters were labeled as Sensitive Areas, Moderate Areas, and Peaceful Areas based on their characteristics, such as average crime rate and types of crimes prevalent in each cluster.

**Cluster Analysis**

* **Sensitive Areas**: Districts in this cluster showed high crime rates, with a significant number of violent crimes such as murders and rapes. Socio-economic factors, ineffective law enforcement, and lower literacy rates were identified as potential contributors to high crime rates in these areas.
* **Moderate Areas**: These districts had average crime rates with a mix of violent and non-violent crimes. Better socio-economic conditions and moderate law enforcement effectiveness were observed in these areas.
* **Peaceful Areas**: Districts in this cluster had low crime rates and fewer incidents of violent crimes. Higher literacy rates, effective law enforcement, and better socio-economic conditions were common characteristics.

**Recommendations**

To reduce crime in sensitive areas, the following actions were recommended:

* **Improving Law Enforcement**: Increasing police presence and improving the efficiency of the legal system.
* **Socio-Economic Development**: Focusing on education, employment opportunities, and poverty alleviation to address root causes of crime.
* **Community Programs**: Implementing community policing and awareness programs to involve citizens in crime prevention efforts.

**Identifying Safe and Unsafe Districts**

The analysis identified the most unsafe districts as those with the highest crime rates and incidents of violent crimes. The safest districts were those with the lowest crime rates and minimal incidents of violent crimes.

**Visualization**

Graphs and visualizations were used to support the findings. A boxplot showed the crime rates by cluster, highlighting the differences between Sensitive Areas, Moderate Areas, and Peaceful Areas. A bar chart illustrated the types of crimes prevalent in sensitive areas, emphasizing the need for targeted interventions.