Crime Hotspot Prediction Using Random Forest and Geospatial Data

Name: Rahul Jaswal

Sap Id: 500107719

Batch 4 (CC&VT)

Submitted To: Dr.Touseef Iqbal

1. Project Objective

The goal of this project is to create a crime hotspot forecasting system that predicts crime-prone locations based on machine learning approaches. The system accepts location, time, and type of crime as input and employs a Random Forest classifier for predicting the occurrence of a crime in certain regions. The answer utilizes Python and Streamlit for building an interactive and user-friendly web interface.

2. Project Workflow

- The process of development is systematic, and the system is made accurate and user-friendly. The project workflow is as follows:
- **Problem Definition**: The problem is to predict crime hotspots by location, time, and type of crime. The system will assist local authorities in identifying high-risk crime areas, allowing for improved resource allocation and crime prevention.
- **Data Collection:** Data is gathered from city crime reports or public crime databases. The data contains important attributes like:
  - Location (district, lat-long)
  - Crime_Type (e.g., theft, assault)
  - Time (hour, day, month)
  - Crime_Frequency (number of incidents)

Data Preprocessing: Missing Values: Replace missing data by filling in or deleting rows.

Timestamp Conversion: Transform time data into appropriate formats (e.g., hour of the day, day of the week).

Encoding: Transform categorical information such as crime type and location into numeric forms for model suitability.

Normalization: Normalize numerical attributes (e.g., frequency of crime) to prepare them for model training.

Feature Engineering: Time-based Features: Derive features such as day of week, time of day, and month to extract temporal trends in crime.

Location-based Features: Geospatial encoding for individual areas or districts based on coordinates or categorization (e.g., urban versus rural).

Crime Frequency: Number of previous crimes within a location and time frame to monitor patterns.

Model Training:

- Random Forest Classifier: A Random Forest classifier is employed to train the model. It utilizes an ensemble of decision trees to make predictions about whether a location will become a crime hotspot.
- Hyperparameter Tuning: Parameters of the model (e.g., number of trees, max depth) are optimized to achieve

maximum prediction accuracy.

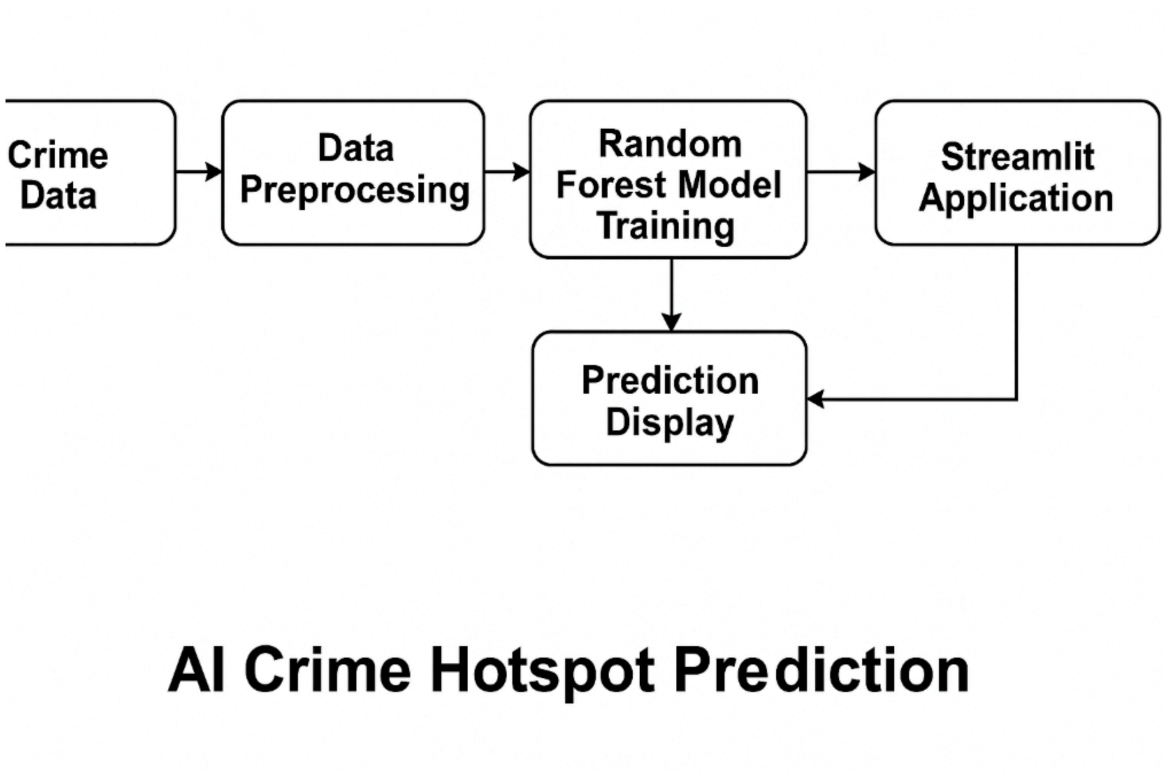
- Prediction: The trained model forecasts the probability of a crime happening at a specific location based on input from users (time, place, type of crime). Predictions are visualized in the form of heatmaps or risk level categories (High, Medium, Low).

Visualization:

- Heatmap: Geographic heatmap to represent crime density over places.
- Bar Charts: A bar chart indicating the risk level of crime for a chosen place.
- Risk Indicator: The model forecasts whether a location is a High-Risk, Medium-Risk, or Low-Risk area for crimes.
- Deployment: The whole system is deployed as a web application using Streamlit, and users can interact with the model in real-time. Users can provide location, time, and crime type, and the system will give back the crime risk prediction.

3. System Design Flowchart

Below is a flowchart representing the system architecture and data flow in the project:



4. Technologies Used

This project is built using the following technologies:

- **Python:** Programming language of choice for implementation.
- **Streamlit:** Library to develop the interactive web interface.
- **scikit-learn:** Machine learning library for the Random Forest Classifier model.
- **Pandas & NumPy:** Numerical calculations and data manipulation.
- **Geopandas/Plotly:** Libraries for geospatial mapping and plots.

5. Machine Learning Model and Features

Random Forest Classifier is employed to predict crime hotspots. It is an ensemble learning algorithm that combines several decision trees to make more accurate predictions. The model is trained on the following attributes:

- **Crime_Type:** The crime type (e.g., theft, assault).
- **Time of Day:** The time of day when the crime was committed.
- **Day of the Week:** Assists in identifying patterns concerning particular days.
- **Location (Latitude & Longitude):** Geospatial attribute that represents the location of where the crime took place.
- **Frequency of Crime:** Number of previous crime occurrences in a specific location and time window.

  These attributes assist the model in learning crime behavior from temporal, spatial, and historic information.

6. Evaluation and Output

The model predicts the crime risk level—high, medium, or low—for a specified location and time based on spatiotemporal and past crime patterns.

- **Visualization:** For the output to be intuitive and actionable, the following visual tools are incorporated within the Streamlit application:

  <u>Crime Risk Heat Map:</u> Interactive map shows predicted risk levels of crime through color gradations. The high-risk areas are represented in red, medium-risk in orange/yellow, and low-risk in green. This interactive map allows users to quickly see where they might need to pay attention or exercise caution.

Prediction Bar Chart: A bar graph shows the model's prediction probability for every level of risk (e.g., 70% high, 20% medium, 10% low). This graph makes users aware of how confident the model is with its prediction.

Risk Level Output Box: A brief, neatly labeled box presents the ultimate classification (e.g., "High Risk Area") so that users can easily see the result without having to look at graphs or data.

User Interaction: The user can provide a location (e.g., in terms of latitude and longitude or area name) and a time (e.g., time of day), and the system dynamically adjusts the prediction and visualizations in real-time.

Output Summary: Area-wise risk distribution interactive heatmap. Bar chart representing prediction score, where likelihoods of each risk level are displayed. Simple text result displaying if the chosen area is low, medium, or high crime risk. This interoperability of textual output and visual analysis brings the system within reach for decision-makers, law enforcement, and the public.

7. Conclusion

This project illustrates a machine learning pipeline to predict hotspots of crime. By incorporating past crime information, feature engineering, and the Random Forest model, this system presents an up-to-date solution to detecting crime-prone hotspots. Local authorities can prioritize prevention of crime and use their resources more effectively through the system.