CSE 3000 - Predictive Crime Modeling

Objective

Develop a crime prediction model using historical crime data, conduct back-testing to evaluate its accuracy, and analyze for potential racial and/or socioeconomic bias in its predictions.

Assignment Details

Part 1: Data Preparation and Algorithm Development

- **Data Collection:**
  - Use publicly available crime data from platforms like:
    - LAPD Open Data
    - Chicago Data Portal
    - NYPD datasets
  - Focus on data with attributes like:
    - **Time:** Date, hour, day of the week.
    - **Location:** Latitude, longitude, precinct/neighborhood.
    - **Crime Type:** Categories like theft, assault, or drug possession.
    - **Demographics:** If available, include race/ethnicity data for offenders/victims.
- **Feature Engineering:**
  - Extract meaningful features (e.g., crime density, time of day, or weather conditions if available).
  - Aggregate spatial data (e.g., divide into grid cells or precincts).
- **Algorithm Training:**

  - Split data into training and test sets (e.g., 80/20).
  - Train a predictive model (e.g., Random Forest, Logistic Regression) to predict:
    * **Crime occurrence in a location (binary classification).**
    * **Crime type in a given area (multiclass classification).**
  - Evaluate accuracy using metrics like precision, recall, F1 score, or confusion matrices.

- **Back-testing:**
  - Use a sliding window or similar technique to test predictions on historical data.
  - Compare predictions to actual crime rates.

Part 2: Bias Analysis

- **Identify Demographic Trends:**
  - Use demographic data of neighborhoods (e.g., census data).
  - Overlay predicted high-crime zones with racial or socioeconomic information.

- **Analyze Prediction Bias:**
  - Evaluate if certain racial or economic groups are disproportionately targeted by high-crime predictions.
  - Quantify bias using statistical methods:
    * **False Positive Rate Disparity:** Measure false positives across demographic groups.
    * **Disparate Impact Analysis:** Calculate if predictions disproportionately affect certain groups.

- **Visualize Results:**
  - Use mapping tools to display predicted hotspots alongside demographic data.
  - Create charts showing differences in error rates or risk scores by demographic group.

Deliverables

- **Code Submission:**
  - Provide a Python script or Jupyter Notebook that implements the tests and analysis.
  - Include well-documented code and instructions.
- **Presentation (5-10 minutes):**
  - Summarize findings using visualizations like maps, charts, and confusion matrices.
  - Propose solutions to mitigate identified biases (e.g., fairer data collection or weighting).
- **Technical Report:**
  - **Data Summary:** Describe the dataset and preprocessing steps.
  - **Algorithm Performance:** Provide back-testing metrics and evaluate model accuracy.
  - **Bias Analysis:** Quantify and discuss any racial and/or socioeconomic disparities.

Example Workflow

- **Predictive Modeling:** Train and evaluate a predictive model using the prepared data.
- **Bias Metrics:**
  - False Positive Rate Disparity.
  - Disparate Impact.
- **Visualization:** Use `Folium` for mapping hotspots and demographics.

Ethical Reflection Prompts

- What are the potential risks of using such models in real-world policing?
- How might biased data (e.g., historical over-policing in minority neighborhoods) skew predictions?
- How could you adjust your algorithm or data collection process to mitigate bias?

Grading Criteria

- **Presentation and Report (40%):** Clarity, visualizations, and technical rigor.
- **Bias Analysis (30%):** Depth of analysis and quality of insights.
- **Algorithm Design (20%):** Data preprocessing, feature engineering, and model accuracy.
- **Ethical Reflection (10%):** Discussion of implications of using this model, and proposed improvements.