Forecasting Fatalities: Insights into US Armed Conflicts in 2023

Authors: Veda Sahaja Bandi, Sindhuja Baikadi

The Issues:

The primary objective of this project is to predict whether fatalities occurred in violent demonstrations in the United States during the year 2023. We aim to achieve this by analyzing past data from 2020 to 2022 and grouping the data by similarity. However, this analysis presents several challenges that require careful consideration:

- How accurately can we predict the occurrence of fatalities in violent demonstrations during 2023 based on historical data from 2020-2022?
- What percentage of locations in 2023 can we correctly identify as having fatalities or not?
- Can our analysis offer meaningful insights into the outcomes of violent demonstrations in the United States?

Addressing these queries is essential for uncovering potential spatial and temporal patterns within the data and setting the stage for further exploration into the dynamics of fatalities in armed conflicts within the United States.

Findings:

This study revealed significant geographical and temporal patterns in armed conflict fatalities. Our analysis achieved an accuracy of 83% in predicting the occurrence of fatalities in violent demonstrations in the United States for the year 2023. This indicates a relatively high level of success in our predictive model based on historical data from 2020-2022.

Despite extensive efforts, our classification method successfully identified only 10 out of 36 fatalities, with 5 accurately predicted fatalities. Conversely, the model accurately classified 176 out of 181 non-fatal incidents. This discrepancy underscores the limitations of our approach in accurately predicting fatalities in violent demonstrations, particularly due to the skewness in the data distribution.

Spatial analysis revealed that certain regions, such as New York, Washington, and California, were identified as hotspots for conflicts. This insight provides valuable information about the geographical distribution of violent demonstrations and can aid in understanding the underlying dynamics driving these conflicts.

These findings highlight the importance of detailed geographical and temporal analysis in understanding and predicting armed conflict patterns in the United States, providing valuable insights for targeted interventions and policy-making.

Discussion:

Our analysis uncovers distinct spatial and temporal trends in armed conflicts throughout the U.S., highlighting regions and periods with heightened conflict fatalities. Despite the model's consistent performance across various parameters, it's important to acknowledge limitations stemming from the exclusive reliance on quantitative data. Further refinement and inclusion of additional variables could enhance predictive accuracy and mitigate misclassifications.

By directing resources towards areas prone to conflict, authorities can implement targeted interventions and policy measures, potentially reducing the impact of armed conflicts and fostering peace and stability. However, the effectiveness of these efforts hinges on a nuanced understanding of the socio-political dynamics underlying these conflicts, which quantitative data alone may not fully capture.

Future research endeavors should consider incorporating more advanced machine learning models and integrating socio-economic and law enforcement variables to obtain a comprehensive understanding of conflict drivers. Continued refinement and exploration of these factors are essential to improving the predictive capabilities of our model and informing policy decisions aimed at promoting peace and security in the United States.

Appendix A: Method

The dataset was obtained in comma-separated values (.csv) file format and was subsequently imported into a Jupyter Notebook for analysis with no null values. The focal point of this analysis was the dataset containing incidents of armed conflicts within the US.

The dataset primarily included geographical coordinates (latitude and longitude) for each incident from the year 2020 to 2023 and whether fatalities occurred or not, which are crucial for spatial and temporal analysis. The data was segregated into training (2020-2022) and testing (2023) sets for model training and evaluation, respectively.

For this project, the K-Nearest Neighbors (KNN) Classification model from the sci-kit-learn Python library was chosen due to its effectiveness in predicting geographical patterns. Its simplicity facilitates easy interpretation and implementation. To determine the optimal number of neighbors for the KNN model, the Elbow Method was employed, plotting the Error Rate against a range of possible odd values of neighbors (1 to 10). The optimal number of neighbors was chosen to balance simplicity with explanatory power, avoiding overfitting or oversimplification.

The model's performance was assessed based on its accuracy in predicting fatalities in the test dataset, with accuracy serving as the primary metric. Additionally, to visualize spatial results, the Folium library was utilized, offering an interactive map displaying armed conflict incidents in 2023.

Overall, the implementation of K-NN Classification coupled with effective data visualization, offered valuable insights into the geographical distribution of armed conflicts in the US during 2023, enhancing our understanding of the dynamics surrounding such incidents.

Appendix B: Results

The dataset comprised 1816 data points representing armed conflicts in the United States from 2020 to 2023, with accompanying latitude and longitude coordinates and information on fatalities. After segregating the data into training (2020-2022) and testing (2023) sets, the K-Nearest Neighbor (KNN) Classifier from the sci-kit-learn library was utilized to analyze geographic patterns in armed conflicts.

The Elbow Method was employed to determine the optimal number of neighbors for the KNN model, with three neighbors (k = 3) identified as providing a reasonable balance between the Error Rate and k values. This optimal parameter selection was crucial for the subsequent modeling process.

Figure 1 illustrates the Elbow Method plot, depicting the trade-off between the number of neighbors and the associated error rate. The plot confirmed the appropriateness of selecting three neighbors for the KNN model based on the dataset characteristics.

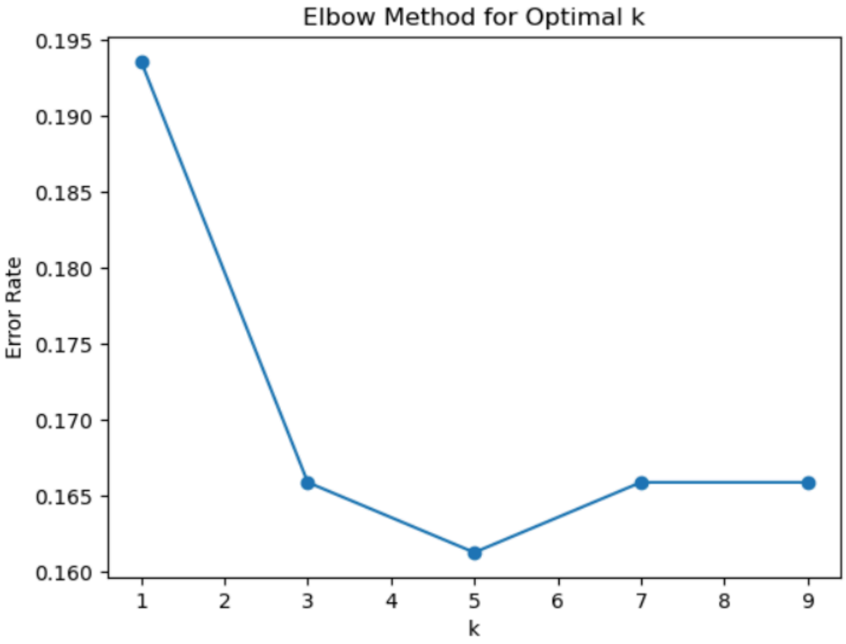

Figure 1: Elbow Method

The KNN Classifier, trained with the optimal number of neighbors, accurately predicted fatalities in 2023 based on data from 2020 to 2022. The resulting visualization, showcased in Figure 2, highlighted fatalities in red and non-fatalities in blue, offering insights into predicted conflict hotspots across the US during 2023.

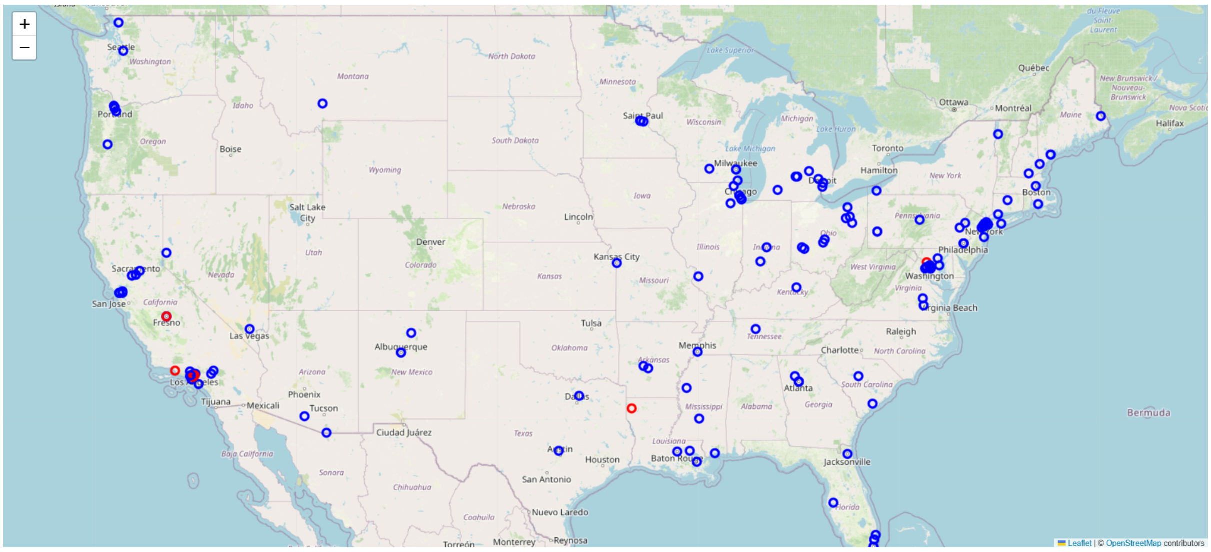

Figure 2: Interactive map with markers

The map visualization, enriched by the color-coded markers, provides a comprehensive depiction of the geographical distribution of incidents across the United States in the year 2023. Notably, there are pronounced concentrations of incidents observed along the East Coast, West Coast, and North Eastern regions. Such geographical patterns hint at potential regional disparities in the frequency of armed conflicts, which could be attributed to diverse factors including population dynamics, socio-political dynamics, levels of urbanization, and variations in policing strategies and practices throughout the nation.

Further analysis of the interactive map revealed pronounced concentrations of incidents along the East Coast, West Coast, and Northeastern regions, particularly in California, New York, and Washington, etc, suggesting potential regional disparities in conflict frequencies. The model demonstrated an 83% accuracy rate in predicting fatalities that occurred in armed conflicts during 2023 in the US.

Due to data skewness, the KNN Classifier identified only 10 out of 36 fatalities, with 5 accurately predicted, while accurately classifying 176 out of 181 non-fatal incidents. We have observed that for odd values of k ranging from 1 to 10, the accuracy of the model consistently hovered around 83% indicating that the model's performance was relatively stable across a range of odd k values.

While the analysis provided valuable spatial insights, it did not directly address the underlying causes of fatalities during armed conflicts. This highlights the need for subsequent investigations into socio-economic, demographic, and law enforcement factors influencing regional and seasonal trends.

In summary, the prediction model offers foundational insights into the spatial distribution of fatal armed conflicts in the US during 2023, laying the groundwork for further investigations and policy formulations aimed at mitigating conflict incidences and their severity.

Appendix C: Code

In this appendix, we document the Python code for performing K-NN Classification to predict the occurrence of armed conflicts in the US during 2023.

```
# Importing the libraries

import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import accuracy_score

from sklearn.metrics import mean_squared_error

import matplotlib.pyplot as plt

import folium

from sklearn.metrics import confusion_matrix

from geopy.distance import geodesic
```

```
# Splitting the data into Training (2020-2022) and Testing (2023) data

train_data = data[data['year'] <= 2022]

test_data = data[data['year'] == 2023]

X_train = train_data[['year', 'month', 'day', 'latitude', 'longitude']]

y_train = train_data['fatalities']

X_test = test_data[['year', 'month', 'day', 'latitude', 'longitude']]

y_test = test_data['fatalities']
```

```
# Elbow Method and Plot

def calculate_error_rate(k):
    knn = KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train, y_train)
    y_pred = knn.predict(X_test)
    return mean_squared_error(y_test, y_pred)

k_values = range(1, 10, 2)

error_rates = [calculate_error_rate(k) for k in k_values]

plt.plot(k_values, error_rates, marker='o')

plt.xlabel('k')

plt.ylabel('Error Rate')

plt.title('Elbow Method for Optimal k')

plt.show()
```

```
# K-NN Classification

knn_3 = KNeighborsClassifier(n_neighbors=3)

knn_3.fit(X_train, y_train)

y_pred = knn_3.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)

cm = confusion_matrix(y_test, y_pred)
```

```
# Metrics to evaluate the model

pd.Series(y_pred).value_counts()

accuracy
```

```
cm
```

Output

```
0   207
1    10

Accuracy: 0.8341013824884793

Confusion Matrix:
[[176   5]
 [ 31   5]]
```

```
# Interactive Map with KNN Classification with predicted values

map_pred = folium.Map(location=[test_data['latitude'].mean(), test_data['longitude'].mean()], zoom_start=5)

for idx, predicted_fatalities in enumerate(y_pred):

    latitude = X_test.iloc[idx]['latitude']

    longitude = X_test.iloc[idx]['longitude']

    color = 'red' if predicted_fatalities == 1 else 'blue'

    folium.CircleMarker([latitude, longitude], color=color, radius=5).add_to(map_pred)

map_pred.save("folium_map_pred.html")

map_pred
```

Contributions:

Sindhuja Baikadi - 02128756: Worked on the Issues, Findings, Discussion, Method, and Results sections. Also self-plotted the graphs to analyze the data using the various methods discussed in the report.

Veda Sahaja Bandi - 02105111: Outlining key observations and insights, worked on the coding portion of the project to implement necessary functionalities and features.