JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY

SECTOR 62, NOIDA, UTTAR PRADESH

INTRODUCTION TO BIG DATA & DATA ANALYSIS

Topic: Global Terrorism Analysis

SUBMITTED BY:

B6 - Shradha Mittal - 22103162

B6 - Dev Agarwal - 22103177

B3 - Swapnil Pandey - 22103061

SUBMITTED TO:

Prof. Sonal Saurabh

Objectives

The objective of this project is to conduct a comprehensive analysis of a global terrorism dataset to derive meaningful insights into the temporal, spatial, and categorical dimensions of terrorist activities. This involves exploring longitudinal trends to understand the progression of attacks over time, analyzing spatial distributions to identify hotspots of activity, and investigating categorical data such as attack types and target types to uncover recurring patterns. Additionally, the project aims to correlate these factors with the scale of casualties, thus providing a multifaceted view of the impact of terrorism. The ultimate goal is to utilize data-driven techniques to better comprehend the dynamics of terrorism and aid stakeholders in decision-making processes. Furthermore, a web-based visualization tool was developed using Streamlit to provide an interactive medium for users to explore the dataset, with a focus on year and country-specific trends.

Problem Statement:

Terrorism is a complex and evolving global issue with widespread social, economic, and political implications. The Global Terrorism Database provides extensive information on terrorist incidents worldwide, but analyzing such a large and complex dataset poses significant challenges. The presence of missing values, inconsistent records, and the multidimensional nature of the data complicates the extraction of meaningful insights.

This project addresses these issues through data cleaning, feature engineering, and robust visualizations to uncover patterns in terrorism. Additionally, a Streamlit-based interface provides an interactive platform for dynamic data exploration, enabling both technical and non-technical users to derive actionable insights.

Methodology

1. Dataset Overview

The global terrorism dataset contains detailed records of terrorist incidents compiled from various sources, encompassing features such as attack type, geographic location, date, weapon type, and casualties. This comprehensive dataset facilitates multidimensional analysis of terrorism-related factors.

2. Data Preprocessing

To ensure analytical accuracy, data preprocessing involved:

- **Handling missing values**: Imputing or removing incomplete entries.
- **Encoding categorical features**: Converting textual categories into numerical formats for model compatibility.
- **Data normalization**: Scaling numerical values to maintain consistency across features.

3. Data Filtering

Data was filtered based on regions, countries, and time periods to extract meaningful subsets. This allowed for:

- Regional and country-specific analysis to identify hotspots.
- Comparative studies of attack types in specific locations or periods.

4. Analysis and Visualizations

To derive insights, various visualizations were used:

- **Line charts**: Illustrating temporal trends in attack frequency.
- **Pie charts**: Showing the proportional distribution of attack types, most active groups and common weapons.
- **Maps**: Highlighting incident locations for spatial pattern recognition.

5. Machine Learning Model

Machine learning models were developed to analyze patterns and predict characteristics of high-impact events.

- **Random forest classifier**: Chosen for its robustness in handling large datasets and uncovering intricate patterns. Trained on historical data, it successfully identified critical factors influencing severe incidents, contributing to predictive insights. The model achieved an impressive 92% accuracy.
- **Decision tree classifier**: Selected for its simplicity and interpretability, the Decision Tree model was employed to visualize decision-making processes based on historical data. By recursively splitting the dataset into subsets, it identified key thresholds and influential variables, providing clear insights into the factors driving high-impact events. The model showed strong predictive capabilities with high interpretability, achieving an accuracy of 89%.

6. Streamlit Dashboard:

- A basic visualization app was developed with filters for country and year.
- Key features include time-series visualizations and maps for attack distribution.

Tools and Technologies Used

- Python
- Jupyter
- Numpy
- Pandas
- Sklearn
- Matplotlib
- Seaborn
- Streamlit

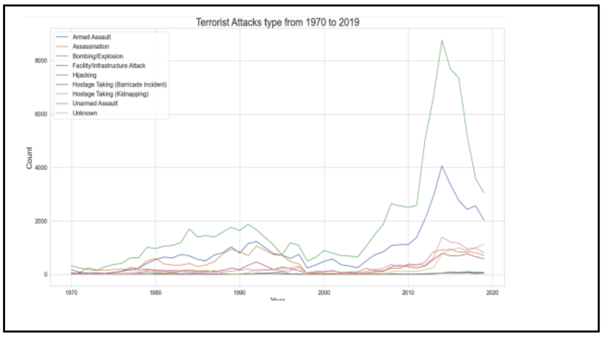
Code Snippets:

1. Data cleaning:

```
df['nkill'] = df['nkill'].fillna(0)

df['nwound'] = df['nwound'].fillna(0)

df['casualty'] = df['nkill'] + df['nwound']
```

2. Visualize Terrorist attack types:

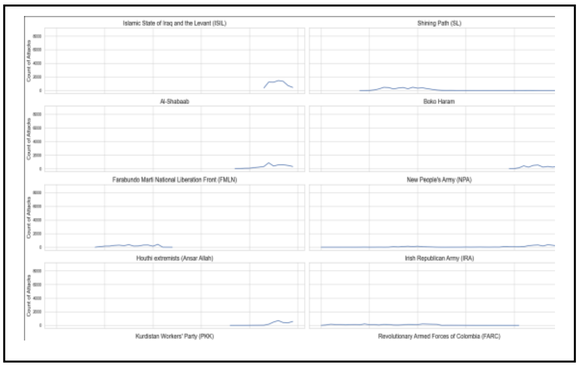
```
plt.figure(figsize=(22, 11))

sns.lineplot(data=df_total_types_killed_year, x='iyear', y='Total_attacks', hue='attacktype1_txt')

plt.title('Terrorist Attacks type from 1970 to 2019', fontsize=25)

plt.legend(fontsize=16)

plt.xticks(fontsize=15)

plt.yticks(fontsize=15)

plt.xlabel('Year', fontsize=20)

plt.ylabel('Count', fontsize=20)
```



3. Most active terrorist groups visualization:

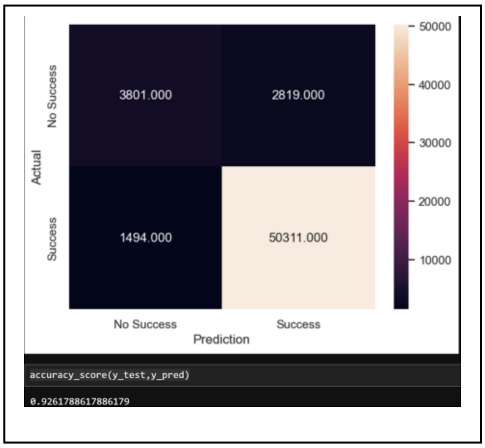
```
# Grouping the data by year and terrorist group

df_total_killed_year_by_group = df.groupby(['iyear', 'gname']).agg(Total_attacks=('nkill', 'count')).reset_index()

# Set up the figure with appropriate size

f, axes = plt.subplots(6, 2, figsize=(22, 18), sharex=True, sharey=True)
```

```
# List of unique terrorist groups to plot
groups_to_plot = df['gname'].value_counts().head(12).index  # Change number to select top N groups
for i, group in enumerate(groups_to_plot):
    ax = axes[i // 2, i % 2]  # Get the appropriate subplot
    sns.lineplot(data=df_total_killed_year_by_group[df_total_killed_year_by_group['gname'] == group],
                 x='iyear', y='Total_attacks', ax=ax)
    ax.set_title(group, fontsize=16)
    ax.set_ylabel('Count of Attacks', fontsize=14)
    ax.tick_params(axis='x', rotation=45)
# Set common x-label and title for the overall plot
plt.xlabel('Year', fontsize=20)
plt.suptitle('Number of Terrorist Attacks by Group from 1970 to 2019', fontsize=25)
plt.tight_layout(rect=[0, 0, 1, 0.96])  # Adjust layout to make room for the title
plt.show()
```



4. Random forest Classifier:

```
from sklearn.ensemble import ExtraTreesClassifier
import matplotlib.pyplot as plt
model = ExtraTreesClassifier()
model.fit(X,Y)
X_train, X_test, y_train, y_test = train_test_split( X, Y, test_size = 0.3, random_state = 100)
classifier = RandomForestClassifier(n_estimators = 10, criterion = "entropy", random_state = 0)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
y_pred
final_model_predictions = pd.DataFrame({'Actual':y_test, 'Predictions':y_pred}).reset_index()
final_model_predictions= final_model_predictions.iloc[:,1:3]
final_model_predictions
```



```
accuracy_score(y_test,y_pred)
0.9261788617886179
```

Results

- Temporal Trends: A significant increase in attacks was observed during certain years, often linked to geopolitical crises.
- Geographic Insights: Regions like the Middle East and South Asia are hotspots, with Iraq and Afghanistan leading in the number of attacks and casualties.
- Attack Types: Bombings and armed assaults dominate, with government and civilian entities as frequent targets.
- Casualty Analysis: Certain attack types and regions are associated with higher casualties.

Challenges

- Data Quality: Missing and inconsistent data required careful handling to ensure analytical integrity.
- Dimensionality: The large number of variables necessitated thoughtful feature selection for meaningful analysis.
- App Optimization: Balancing simplicity and functionality in the Streamlit app was critical to delivering an intuitive user experience.

Conclusion

In conclusion, this project seeks to provide a comprehensive analysis of terrorist activities by examining global and regional trends over time. By identifying the factors linked to high-impact events, including the types of attacks, target groups, and weapons used, we aim to uncover critical insights that can inform more effective counter-terrorism strategies. Furthermore, leveraging machine learning techniques to explore and analyze data enhances our predictive capabilities, allowing us to anticipate future incidents and potentially mitigate their impact. Ultimately, this project aspires to contribute valuable knowledge to the field of terrorism studies, supporting efforts to enhance global security and promote resilience in the face of evolving threats.