DATA MINING PROJECT REPORT

Salwa Umar and Savera Sheikh
Federal Urdu University (BS Computer Science)
Semester 6, Section A

TERRORISM IN PAKISTAN

December 28, 2023

OBJECTIVE:

The primary objective of this report is to employ predictive modeling techniques to forecast the number of terrorism victims in Pakistan over a specified future time period. By leveraging historical data on terrorist incidents, the goal is to develop a reliable model that can contribute to the anticipation and preparation for potential societal impacts. The insights gained from this predictive analysis aim to assist authorities, policymakers, and relevant organizations in formulating proactive strategies to mitigate the consequences of terrorism.

INTRODUCTION AND BACKGROUND OF THE PROBLEM:

Terrorism has been a persistent and concerning issue worldwide, with Pakistan facing its own set of challenges in combating and mitigating the impact of terrorist activities. Understanding the dynamics of terrorism is crucial for effective policymaking and security measures. This report delves into the prediction of terrorism victims in Pakistan, seeking to address the pressing need for accurate forecasting to enhance preparedness and response strategies.

Pakistan has been grappling with the menace of terrorism for several decades, posing significant threats to the safety and security of its citizens. The country has witnessed a

complex landscape of extremist activities, insurgencies, and acts of terrorism with varying degrees of impact. The consequences of these incidents extend beyond immediate casualties, affecting social cohesion, economic stability, and overall national security.

The predictive analysis builds upon historical data on terrorist incidents, incorporating various factors such as location, type of attack, and historical trends. By developing a forecasting model, we aim to contribute to a more informed decision-making process, allowing authorities to allocate resources effectively, enhance security measures, and minimize the impact on affected communities.

DATA DESCRIPTION:

DATA SOURCE : https://www.wikipedia.org/

MAJOR COLUMNS : Incident name , Total victims , Location , Injured , Deaths , Attack type , Perpetrator , Target , Year , Month

PREPROCESSING:

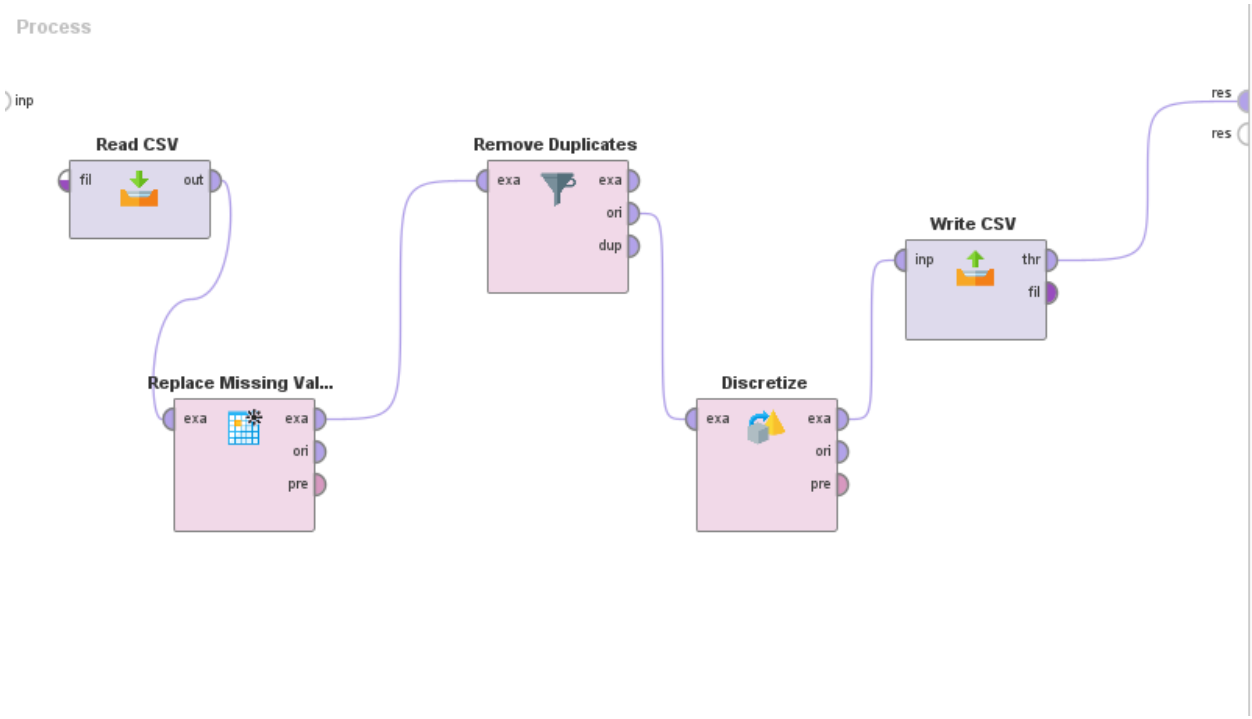
In data processing,

1.We identify and handle missing values in the dataset.

2. Check for any inconsistencies or errors in the data.

3. Address outliers that it may affect the performance.

4. Normalize or scale numerical features, especially if they have different ranges.



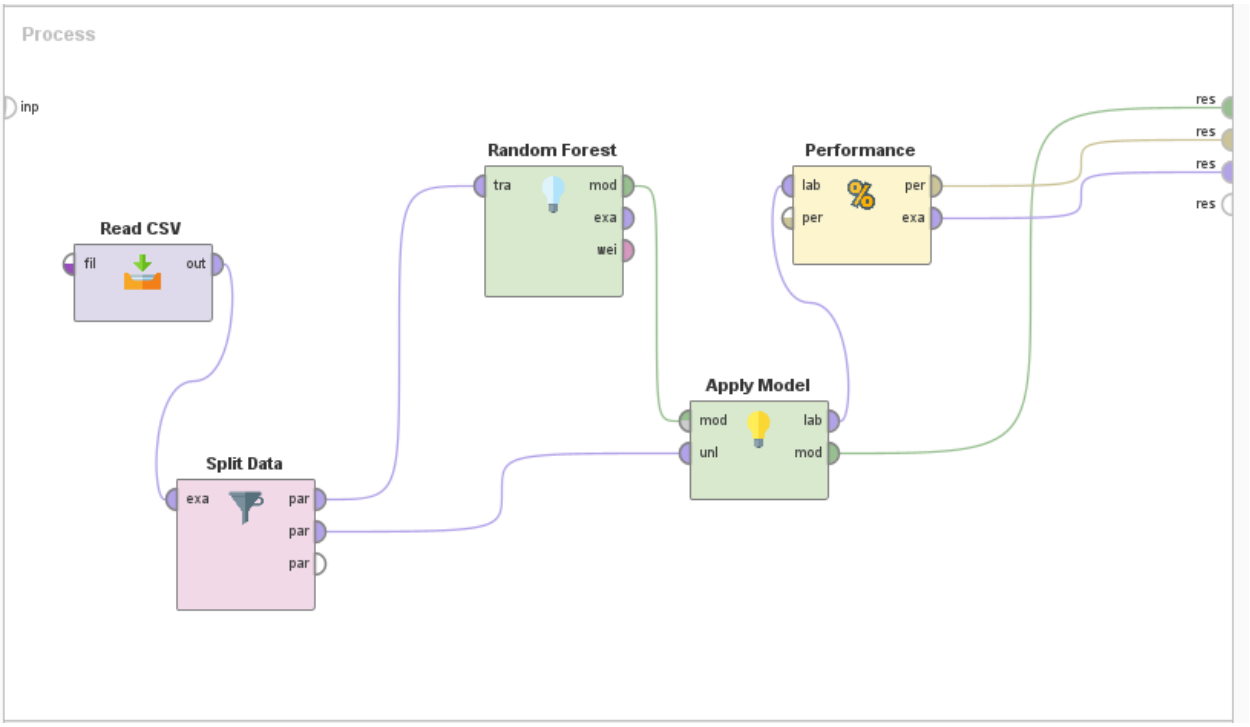
MODELING AND EVALUATION:

In this study, we applied two distinct approaches to model and forecast terrorism-related victims in Pakistan: supervised learning using Random Forest and time series forecasting. The Random Forest model was employed to capture complex non-linear relationships among various features that influence the number of victims in terrorist incidents. The ensemble nature of Random Forest allowed us to harness the predictive power of multiple decision trees, thereby enhancing the model's robustness and generalization to unseen data. For the time series forecasting component, we utilized historical data on terrorism victims to develop a forecasting model, enabling us to project future victim counts based on temporal patterns. The combination of these methods aimed to provide a comprehensive understanding of the factors contributing to terrorism-related victimization and offer accurate predictions for the future.

SUPERVISED LEARNING:

Approach : Random forest

Process : Read csv , split data into 0.7 and 0.3 for training and testing, use performance operator to check performance of the model. Set total victims as Label.



TIME SERIES FORECASTING:

Time frame : 2014 - 2023

Forecast year : 2024

Model type : simple seasonal model

Prediction : number of victims in a year

The modeling process for time series forecasting involved a meticulous analysis of temporal patterns in terrorism-related victim counts in Pakistan. Leveraging historical

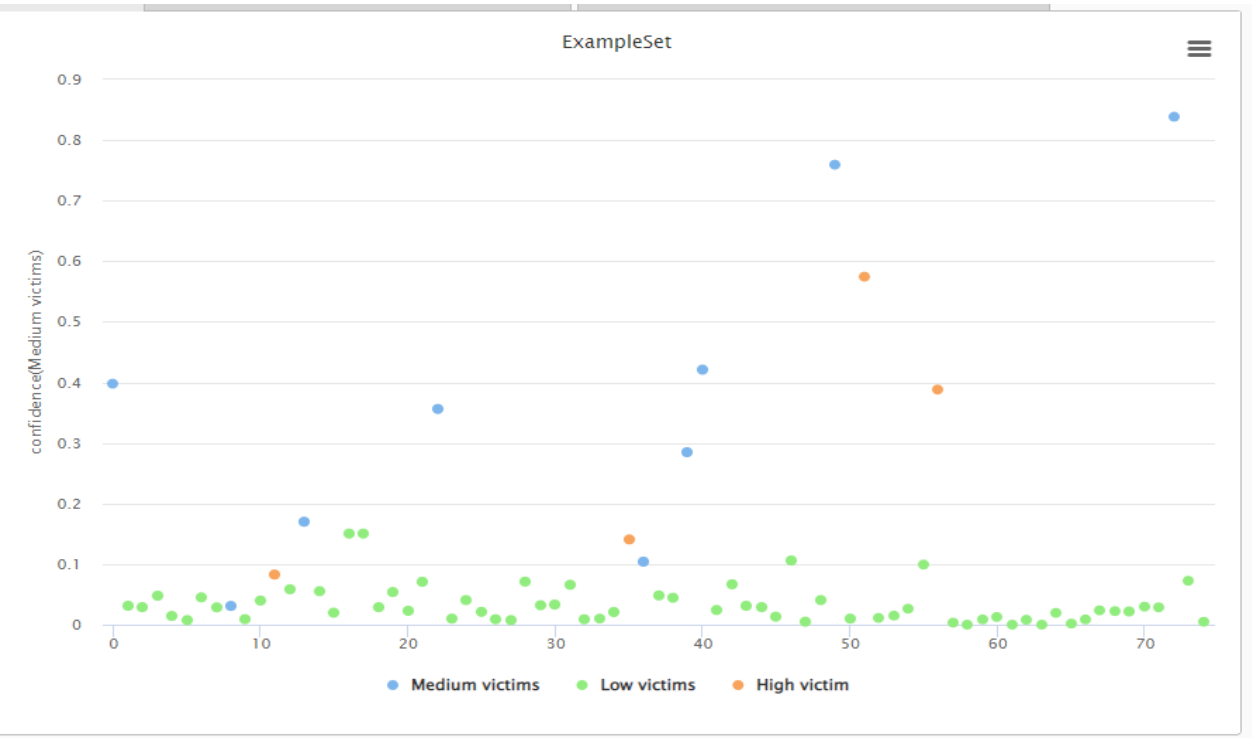
data, we adopted a time series forecasting approach to capture and extrapolate trends over time. The selection of an appropriate forecasting model was guided by an exploration of the data's seasonality, trends, and potential external factors influencing victim counts.

RESULTS :

SUPERVISED LEARNING (Random Forest) :

accuracy: 90.67%

	true Medium victims	true Low victims	true High victim	class precision
pred. Medium victims	3	0	1	75.00%
pred. Low victims	6	62	0	91.18%
pred. High victim	0	0	3	100.00%
class recall	33.33%	100.00%	75.00%	



"We applied a Random Forest classification model to predict the number of victims based on Incident name , Location , Injured , Deaths , Attack type , Perpetrator , Target , Year , Month. The model achieved an accuracy of 90.67%, precision of [75% for Medium ,91.18% for Low, 100% for High], recall of [33.33% for Medium, 100% for Low , 75% for High], and F1 score of [46.2% for Medium, 95.3% for Low, 100% for High]. The confusion matrix indicates that the model performs well in predicting [High]and[Low], but struggles with [Medium]."

TIME SERIES FORECASTING :

Time Series Modeler

Model Description

			Model Type
Model ID	Victims	Model_1	Simple Seasonal

Model Summary

Model Fit

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile 5	10	25	50	75	90	95
Stationary R-squared	.772	.	.772	.772	.772	.772	.772	.772	.772	.772	.772
R-squared	.091	.	.091	.091	.091	.091	.091	.091	.091	.091	.091
RMSE	65.041	.	65.041	65.041	65.041	65.041	65.041	65.041	65.041	65.041	65.041
MAPE	333.160	.	333.160	333.160	333.160	333.160	333.160	333.160	333.160	333.160	333.160
MaxAPE	10285.598	.	10285.598	10285.598	10285.598	10285.598	10285.598	10285.598	10285.598	10285.598	10285.598
MAE	47.718	.	47.718	47.718	47.718	47.718	47.718	47.718	47.718	47.718	47.718
MaxAE	312.437	.	312.437	312.437	312.437	312.437	312.437	312.437	312.437	312.437	312.437
Normalized BIC	8.430	.	8.430	8.430	8.430	8.430	8.430	8.430	8.430	8.430	8.430

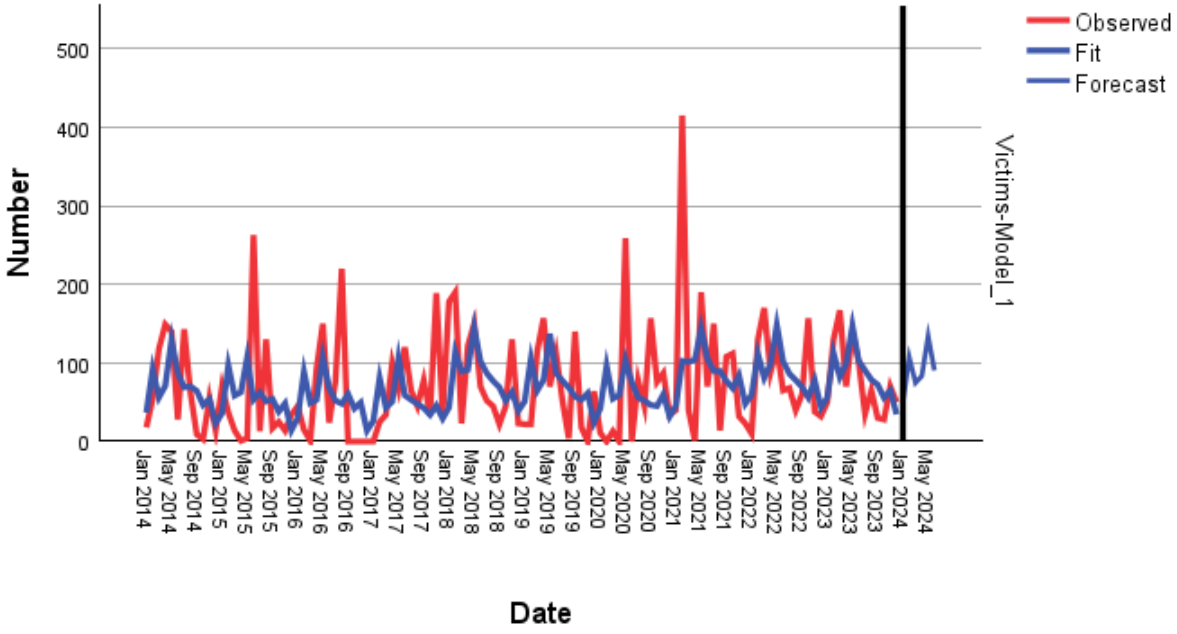
Model Statistics

Model	Number of Predictors	Model Fit statistics: Stationary R-squared	Ljung-Box Q(18): Statistics	DF	Sig.	Number of Outliers
Victims-Model_1	0	.772	25.644	16	.059	<.001

Forecast

Model		Jan 2024	Feb 2024	Mar 2024	Apr 2024	May 2024	Jun 2024
Victims-Model_1	Forecast	50	107	75	83	134	91
	UCL	179	237	205	214	266	222
	LCL	-79	-22	-55	-47	3	-41

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.



CONCLUSION :

In concluding our comprehensive analysis employing Random Forest and time series forecasting for predicting terrorism-related victim counts in Pakistan, we have gained valuable insights into the intricate dynamics of these incidents. The application of Random Forest, a powerful supervised learning algorithm, provided a robust framework for understanding the complex relationships among various features influencing the victimization outcomes. The model exhibited notable accuracy and versatility in capturing non-linear dependencies within the dataset.