MACHINE LEARNING PROJECT

S1

NAME: HIFZA MOIN

COURSE: MACHINE LEARING 1

ERP ID: 23662

PROJECT NAME: WORLDWIDE TERRORISM

- **S2): Corporate domain of your project (finance, retail, etc.)**

My project is about how terrorism spreads in world, database of global terrorism. The Global Terrorism Database documents more than 200,000 international and domestic terrorist attacks that occurred worldwide since 1970. It covers many aspects from terrorist groups to the country, region, total number of people killed, weapons used in attacks and many other aspects. This data has been used worldwide for prediction and analysis of terrorism, which country is on at risk, different countries are using this database for prediction and condition of their own country, to see the pattern, to learn more about terrorist groups and danger entitled with that group.

- **(S3): Background Knowledge about this domain and about the data you have:**

The Global Terrorism Database (GTD) documents more than 200,000 international and domestic terrorist attacks that occurred worldwide since 1970.

It includes information about

1. Includes information on more than 95,000 bombings, 20,000 assassinations, and 15,000 kidnappings and hostage events since 1970.
2. It holds information about how many peoples have been killed by terrorist, their nationality, their country, their region and many other details.
3. A large variety of information about the weapons and explosives used in attacks, types of weapon and sub types of weapons for better understanding of attacks.
4. A brief detail about the group involved in the attacks, what were their real targets domain.
5. Details about the nature of attack, either it was suicidal or not, was it successful or not, who was the bomber, is the host kid, the latitude and longitude of the attacked country.

6. The details about Global distribution of terrorism, regions experience the most terrorism, Share of deaths from terrorism by country.
7. Incremental growth of terrorism over years and many other aspects.

- **(S4): Problems:**

In my project I basically worked on 3 different aspects of this data. The data contains high numbers of missing values so I wrangled the data and minimize the features as much as possible depends on the scenario I am working, the data itself contains a total of 136 columns, which I minimized to 10 to 15 to train my model.

I covered 3 aspects of this problem as I highlighted in my background knowledge.

- Predicting the success of attacks, was the attack successful or not.
- Predicting the name of group that carried out the attack.
- Predicting the attacks based on its type, I worked on suicidal attack, because most of the terrorist attacks were the suicide bomb explosive attacks.

As their still can a lot to be done with this data, some of the problems can be done with data are

- Time series forecasting, as the data evolves and increased over time so we can predict the place of attack or the loss by attacks.
- Ml problem that can predict the place, country or region of attack.

- **(S5): Pipeline:**
    1. Import all the libraries.
    2. Take a deep insight and understands the data, shape, information, values, columns and others.
    3. Finding the patterns in data, work with missing values and outliers.
    4. Dropping the irrelevant columns and columns whose most of the entries are missing.
    5. Impute the other missing values by using different functions.
    6. Visualizing the data by using different graphs and plots.
    7. Use encoding for categorical variables.

8. Convert data into parts train and test.
9. Used feature selection algorithm to take relevant features.
10. Used correlation heatmaps, t-tests, Anova and other algorithms to understand the relationship with my label class.
11. Class imbalancing.
12. Scaling
13. Use different models to train and predict my data.
14. Check prediction, if it's working correctly

- **(S6): ML-Projects:**

All the wrangling and other project requirements are in Notebook along with an explanation.

A I mentioned before I worked on 3 different aspects

- For predicting success: I am using fscore as my measure because I care more about the positive class.
- For predicting Groups involved: I'm using accuracy as my measure because my data is balanced and every class is important to me.
- For predicting about suicidal attacks: My measure is this AUC and accuracy, because I care equally about positive and negative classes.

I've used different feature selection algorithm for my different prediction, I used random forest classifies, extra tree classifier, f_classif, Anova, t-tests and correlation maps.

1. **What effect is cross-validation having on my results?**

It helps avoid overfitting in my data and estimate the skill of the model on new data. It helps determining the parameters of my model, in the sense that which parameters will result in the lowest test error.

2.What effect is feature selection having on my results?

It helps in selecting those features in data that are most useful or most relevant for the problem you are working on, as I had many features like 136 after eliminating many because of nan values I used feature selection to reduce dimensionality because it won't be as accurate if I train my model on many variables instead, I chose the more relevant related to my problem.

3. Which combination is best? (Based on test data performance)

K-fold cross validation with random forest algorithm.

Now tell how your results are solving the problem you identified above?

My result will solve the problem by using feature selection algorithms, k-fold validation, and trained my data on machine learning algorithms and chose the best by using different measures.

Suppose your model is deployed in industry and generates predictions on live test data. What if the distribution of the original data changes, e.g., the customers' behavior pattern changes? Then your model will start to give the wrong predictions. How will you update your model in this situation?

By training my model on current data to adapt the changes.