**Global Terrorist Attack - Challenge**

**Dataset Consider**:

1) gtd\_11to14\_0615dist (2011 to 2014)

2) gtd\_92to10\_0615dist (1992 to 2010)

**Dictionary for Terrorist GroupName :** Generate Mapping of Terrorist Group Name to Unique Id and prepared dictionary for it.

I found total **1892** different groups of terrorist attack from dataset.

**Data PreProcessing Task:**

Handle missing value data

Remove inconsistent data

If two or more group claimed for same attack on same day then i consider these all groups as separate input raw with its claimed result.

**Feature Selection:**

In Data Set i found out different 134 features and selected following relevant features:

List of Features:

1. Month
2. Day
3. Country
4. Attacktsuccess
5. Attacktype
6. Targtype
7. GroupName
8. Weaptype
9. Claimed

**Machine Learning Tool used** : Weka

Filter : NumericToNominal

No of attack did not claimed by group : 63736

No of attack claimed by group: 10331

**Machine Learning Algorithm Selection :**

After feature selections task i found out most of the features are totally independent to each other.

So Naive bayes and other probabilistic algorithm are most relevant for this kind of problems.

Applied various probabilistic machine learning algorithms like Naive Bayes, Random Forest and ID3 to solve this Challenge and found best accuracy using Random Forest algorithm.

**Best Accuracy Details :**

Random Forest Algorithm :

5 Fold cross Validations

=== Classifier model (full training set) ===

Random forest of 30 trees, each constructed while considering 4 random features.

Out of bag error: 0.0838

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 67822 91.5684 %

Incorrectly Classified Instances 6245 8.4316 %

Kappa statistic 0.6156

Mean absolute error 0.1147

Root mean squared error 0.2466

Relative absolute error 47.7902 %

Root relative squared error 71.1665 %

Total Number of Instances 74067

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.968 0.405 0.936 0.968 0.952 0.943 0

0.595 0.032 0.749 0.595 0.663 0.943 1

Weighted Avg. 0.916 0.353 0.91 0.916 0.912 0.943

F-Measure = 91.20%

=== Confusion Matrix ===

a b <-- classified as

61676 2060 | a = 0

4185 6146 | b = 1

**Naive Bayes Algorithm :**

5 Fold cross Validations

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 66475 89.7498 %

Incorrectly Classified Instances 7592 10.2502 %

Kappa statistic 0.562

Mean absolute error 0.1288

Root mean squared error 0.267

Relative absolute error 53.6481 %

Root relative squared error 77.0625 %

Total Number of Instances 74067

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.945 0.397 0.936 0.945 0.941 0.933 0

0.603 0.055 0.641 0.603 0.621 0.933 1

Weighted Avg. 0.897 0.35 0.895 0.897 0.896 0.933

F-Measure = 89.6%

=== Confusion Matrix ===

a b <-- classified as

60249 3487 | a = 0

4105 6226 | b = 1

Here I found best accuracy for Random forest Algorithm but in future we can use more algorithms then take predictions as majority predictions from different machine learning algorithms and consider the result based on Majority Voting System.