RV COLLEGE OF ENGINEERING®

(Autonomous Institution Affiliated to VTU, Belagavi)
Accredited by National Board of Accreditation, New Delhi
R.V. Vidyanikethan, Bengaluru-560059

DEPARTMENT OF MASTER OF COMPUTER APPLICATIONS

Assignment

III Semester

Machine Learning(18MCA343)

Submitted by

USN	Name of the Candidate
1RZ18MCA01	Vivin Adhitya A M
1RZ18MCA14	Harshith Kumar K
1RZ18MCA15	Hemanth Varma B H

December 2019

RV COLLEGE OF ENGINEERING®

(Autonomous Institution Affiliated to VTU, Belagavi)
Accredited by National Board of Accreditation, New Delhi
R.V. Vidyanikethan, Bengaluru-560059

DEPARTMENT OF MASTER OF COMPUTER APPLICATIONS

CERTIFICATE

This is to certify that the following students have successfully completed the Assignment on **Machine Learning (18MCA343)** in a partial fulfillment of III semester MCA during the academic year 2019-20.

USN	Name of the Candidate
1RZ18MCA01	Vivin Adhitya A M
1RA18MCA14	Harshith Kumar K
1RZ18MCA15	Hemanth Varma B H

USN	Name of the Candidate	Marks	
		Max	Obtained
1RZ18MCA01	Vivin Adhitya A M	30	
1RA18MCA14	Harshith Kumar K	30	
1RZ18MCA15	Hemanth Varma B H	30	

Dr. Andhe Dharani
Professor and Director
Department of MCA
R.V.College of Engineering
Bengaluru - 560059

Dr. Andhe Dharani
Professor and Director
Department of MCA
R.V. College of engineering
Bengaluru - 560059

Table of contents

CHAPTER 1

Title - Global Terrorism analysis and prediction

ABSTRACT

Terrorism is the unlawful use of force or violence against persons or property to intimidate or coerce a government, the civilian population, or any segment thereof, in furtherance of political or social objectives. It targets ethnic or religious groups, governments and political parties, corporations and media enterprises. Terrorism that occurs throughout the world is known as global terrorism. It is probably the worst type of crime that ever exists. Not only does it kill people, it destroys livelihoods, economies, and civilized world order that took millennia to form. The results of terrorism are almost always catastrophic. Individuals or groups that commit these crimes are called terrorists. Terrorists exist all over the world. There are a few that operate alone, but mostly they are parts of one of many global organizations.

INTRODUCTION

Terrorism affects people individually whether it is aimed directly to them or to people around them. As a result, people can lose their lives, family and livelihood. Terrorism affects society and families in a major way. When a terrorist attacks or destroys a building, public transport, houses etc. it affects the whole community. It is not only affecting the community by destroying something, but it also takes lives and family members. Terrorism has almost become a part of the modern society. In many parts of the world Separatists, religious fanatics and mentally abnormal had have try to challenge established Governments and have released waves of murder, violence and terror. Most of the day we come across violence, explosion, landmine blast or other aids of terrorism. The U.S. has suffered a lot of its casualties because of suicide bombings in recent years. The government has stepped up its security to try and stop these disasters, but it has been difficult. There are many other organizations and groups that use terror for other purposes besides religion. Terrorism dates back to at least the 1st century, when the Zealots, a Jewish religious sect, fought against Roman occupation of what is now Israel. At the beginning of the 19th century, terrorist

movements acquired a more political and revolutionary direction. In the late 19th and early 20th centuries, anarchists in Italy, Spain, and France used terroris.

Motivation

We wished to use this project as an opportunity to shed light on some serious issues we are facing globally. Among these were issues like animal rights, environmental threats, human rights, and terrorism. According to a survey, about 218 million people are affected by calamities, natural and man-made, per annum and about 68000 people lose their lives every year. The frequency of natural disasters like earthquakes, volcanoes, etc. have remained broadly constant, but the number of terrorist activities have grown over the period. Thus terrorism seemed like an appropriate issue to bring up. Also, we tried to look for data sources for all those issues and found readily available detailed data for global terrorism that would supplement our analysis.

Objective

The purpose of this project is to predict the success of the terrorist attack given a set of input features.The Global Terrorism Database (GTD) documents more than 190,000 international and domestic terrorist attacks that occurred worldwide since 1970. With details on various dimensions of each attack, the GTD familiarizes analysts, policymakers, scholars, and journalists with patterns of terrorism. The GTD defines terrorist attacks as: Acts by non-state actors involving the threatened or actual use of illegal force or violence to attain a political, economic, religious, or social goal through fear, coercion, or intimidation. Data collection is ongoing and updates are published annually at www.start.umd.edu/gtd.The database sourced by unclassified media articles contains information on multiple dimensions of each event. More than 100 structured variables characterize each attack's location, tactics and weapons, targets, perpetrators, casualties and consequences, and general information such as definitional criteria and links between coordinated attacks.

CHAPTER 2

LITERATURE SURVEY

Sl No	Title of the Paper with Authors	Details of Publication	Summary of the Paper	References
01	Global Terrorism Abadie, Alberto.	Fortna, Virginia Page (May 20, 2015)	"Do Terrorists Win? Rebels' Use of Terrorism and Civil War Outcomes".	*International Organization*. **69** (3): 519–556. doi:10.1017/S0020818315000089. hdl:1811/52898.
02	Terrorist Attacks against Religious Targets in the United States. **Erin Miller** .	The National Consortium for the Study of Terrorism and Responses to Terrorism .	Between 1970 and 2017, 150 terrorist attacks in the United States targeted religious figures and institutions. Fifteen of these attacks were lethal.	https://www.start.umd.edu/sites/default/files/publications/local_attachments/START_ReligiousTargets_FactSheet_Oct2018.pdf
03	Ideological Motivations of Terrorism in the United States.	Miller, Erin. 2017. "Ideological Motivations of Terrorism in the United States,	Terrorism is a narrowly defined type of violence, even within the broader spectrum of ideologically	https://www.start.umd.edu/publication/ideological-motivations-terrorism-united-states-

		1970-2016." College Park, Maryland. November.	motivated violence.	1970-2016
04	Why Terrorism Does Not Work. International Security 31. Abrahms, Max	2018 Impact Factor: 4.500 2018 Google Scholar h5-index: 33	This is the first article to analyze a large sample of terrorist groups in terms of their policy effectiveness. It includes every foreign terrorist organization (FTO) designated by the U.S. Department of State since 2001	https://www.mitpressjournals.org/doi/10.1162/isec.2006.31.2.42
05	Live to Win Another Day: Why Many Militant Organizations Survive yet Few Succeed. Acosta, Benjamin. 2014.	journal **Studies in Conflict & Terrorism** Volume 37, 2014 - Issue 2	Militant organizations pursue two common aims: to survive and to achieve the goals.	https://www.tandfonline.com/doi/abs/10.1080/1057610X.2014.862900

Hardware requirements

- Quad Core CPU (multiple recommended)
- 4 GB of RAM or higher recommended

- Disk Space Storage Requirements(10GB)
- Redundant power

Methods of Data Collection-

1 We used datasets from Kaggle - Results of the attacks since 1970 and 2018.

2 Data from https://start.umd.edu/

3 Global Terrorism Database

Tools Used

Spyder

Spyder is a powerful scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It offers a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package.

Libraries used

Numpy

NumPy is a Python package which stands for 'Numerical Python'. It is the core library for scientific computing, which contains a powerful n-dimensional array object. NumPy (Numerical Python) is a linear algebra library in Python. It is a very important library on which almost every data science or machine learning Python packages such as SciPy (Scientific Python), Mat-plotlib (plotting library), Scikit-learn, etc depends on to a reasonable extent. NumPy is very useful for performing mathematical and logical operations on Arrays. It provides an abundance of useful features for operations on n-arrays and matrices in Python.

Pandas

pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. Pandas is the most popular python library that is used for data analysis. It provides highly optimized performance with back-end source code is purely written in *C* or *Python*.

Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. One of the best but also more challenging ways to get your insights across is to visualize them that way, you can more easily identify patterns, grasp difficult concepts or draw attention to key elements. When you’re using Python for data science, you’ll most probably will have already used Matplotlib, a 2D plotting library that allows you to create publication-quality figures. Another complimentary package that is based on this data visualization library is Seaborn, which provides a high-level interface to draw statistical graphics.

Matplotlib

Matplotlib pyplot is a plotting library used for 2D or 3D graphics in python programming language. At first sight, it will seem that there are quite some components to consider when you start plotting with this Python data visualization library. You’ll probably agree with me that it’s confusing and sometimes even discouraging seeing the amount of code that is necessary for some plots, not knowing where to start yourself and which components you should use.

Scikit-learn

Scikit-learn is probably the most useful library for machine learning in Python. It is on Numy, SciPy and matplotlib, this library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression,clustering and dimensionality reduction.

CHAPTER 3

Implementation

Algorithm used - Gaussian Naive Bayes

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

Main Reason for using random forest algorithm

- If the independence assumption holds then it works more efficiently than other algorithms.
- It requires less training data.
- It is highly scalable
- It can make probabilistic predictions.
- Can handle both continuous and discrete data.
- It can work easily with missing values.
- It is fast and can be used to make real-time predictions
- It doesn’t require as much training data

Importing Data

Spyder (Python 3.7)

File Edit Search Source Run Debug Consoles Projects Tools View Help

C:\Users\harsh

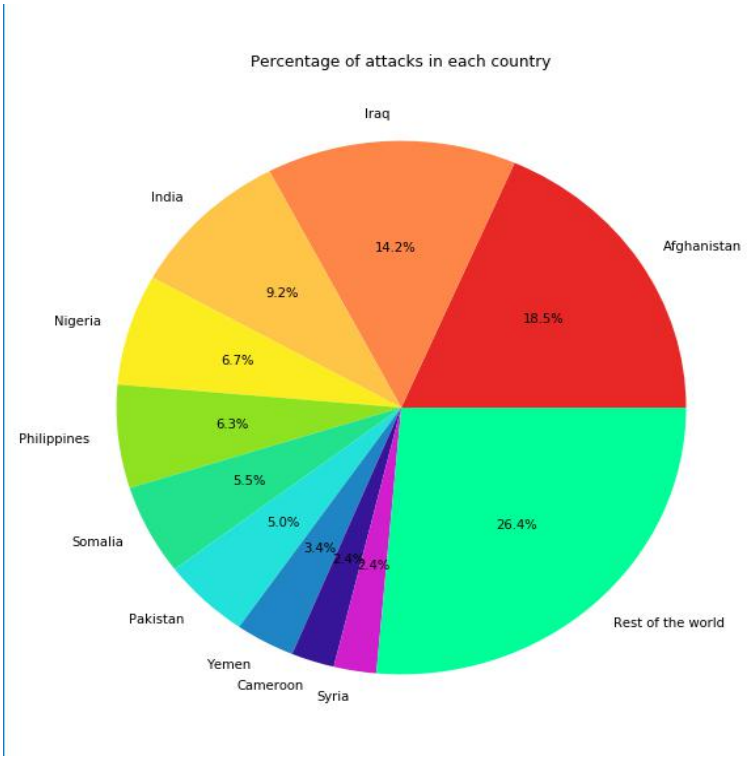
Editor - C:\Users\harsh\.spyder-py3\temp.py

temp.py | MLFirstPhase.py | ML2.py - Machine Learning | india.py | cle.py | ML2.py - Desktop

```
1 import numpy as np
2 import pandas as pd
3 import matplotlib.pyplot as plt
4 import itertools
5 import csv
6 from sklearn.naive_bayes import GaussianNB
7 from sklearn.metrics import accuracy_score, f1_score
8 from sklearn.model_selection import train_test_split
9 import seaborn as sns
10 from sklearn.metrics import confusion_matrix
11 df = pd.read_excel(r'C:\Users\harsh\OneDrive\Desktop\globalterrorism.xlsx', encoding='ISO-8859-1')
12
```

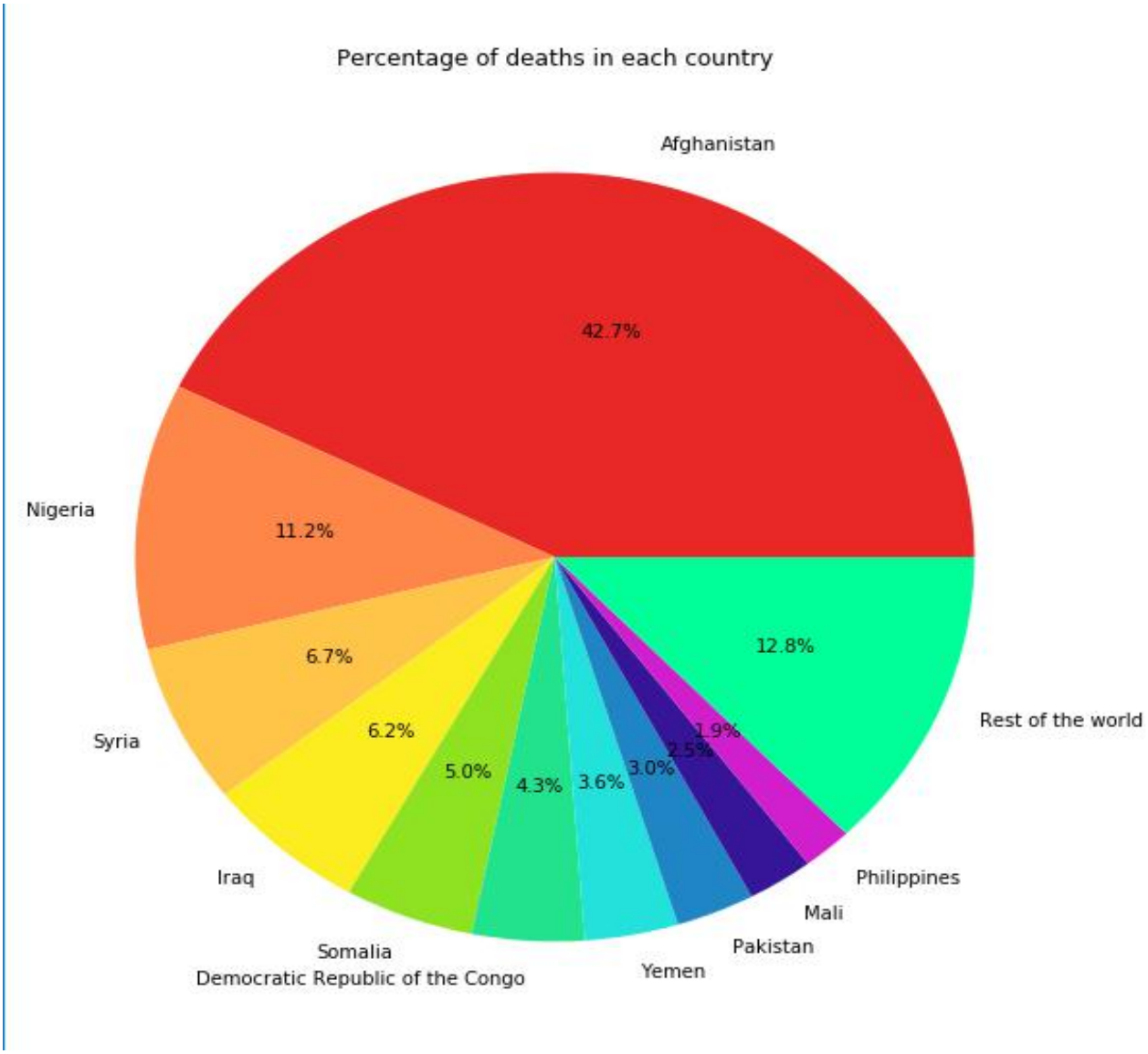
Percentage of attacks in each country

For year 2018, more than 75% of terrorist attacks happened within just 10 countries : Iraq, Afghanistan, India, Somalia, Turky, Yemen, Nigeria, Syria, Philippines and Pakistan. Rest of the world combined had less than 25% of total global terrorist attacks.



Percentage of death in each country

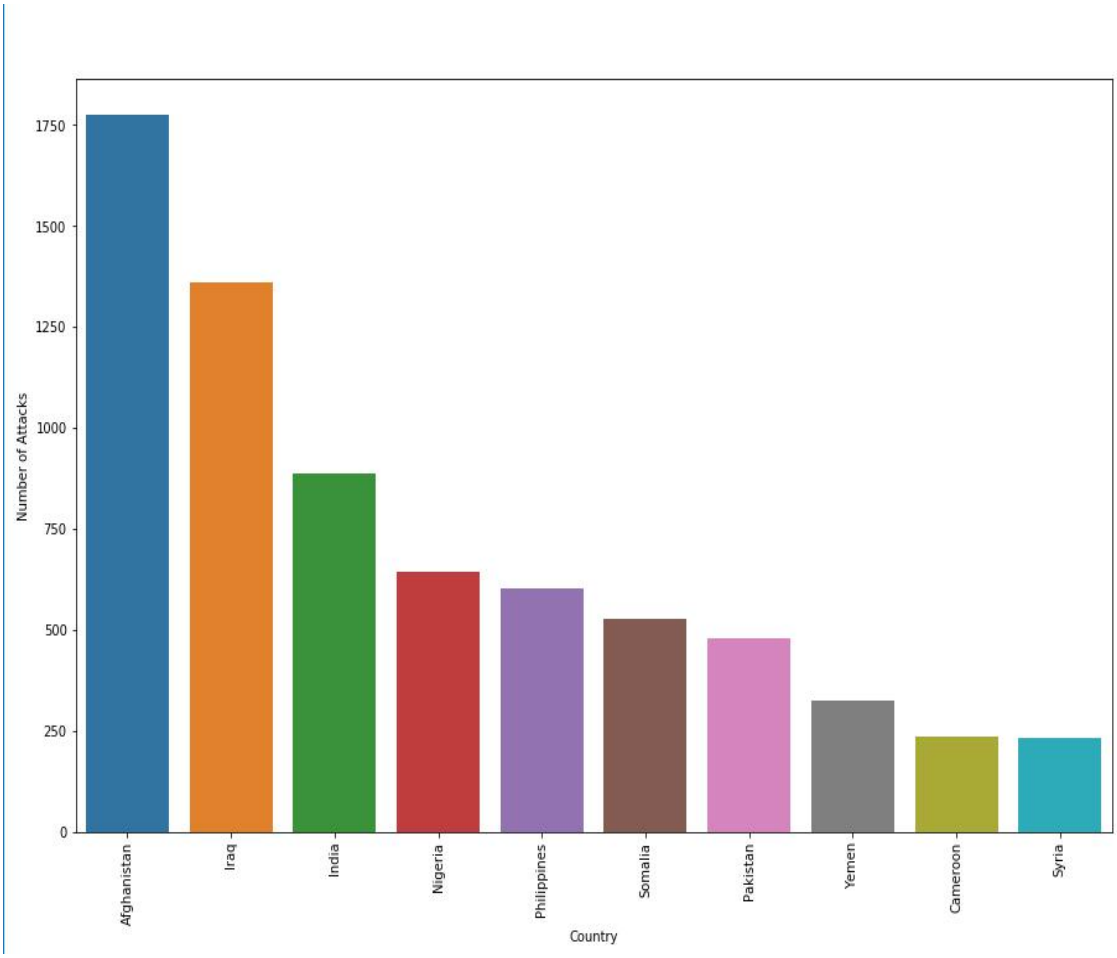
For year 2016, more than 86% of deaths due to terrorist attacks occurred within 10 countries : Iraq, Afghanistan, Syria, Nigeria, Somalia, Yemen, Turkey, Pakistan, South Sudan, Libya. Rest of the world combined had less than 14% of deaths.



In the above graph we can know that Afghanistan is the country which is affected more by terrorist attack with 42.7% when compared to rest of the world

Total of attacks in each country in the year 2018

Now below is the bar graph of number of attacks for top 10 countries for year 2016. Looking at the graph it seems that Iraq is the most dangerous country to live in, then comes Afghanistan, India, Somalia, Turkey and rest.



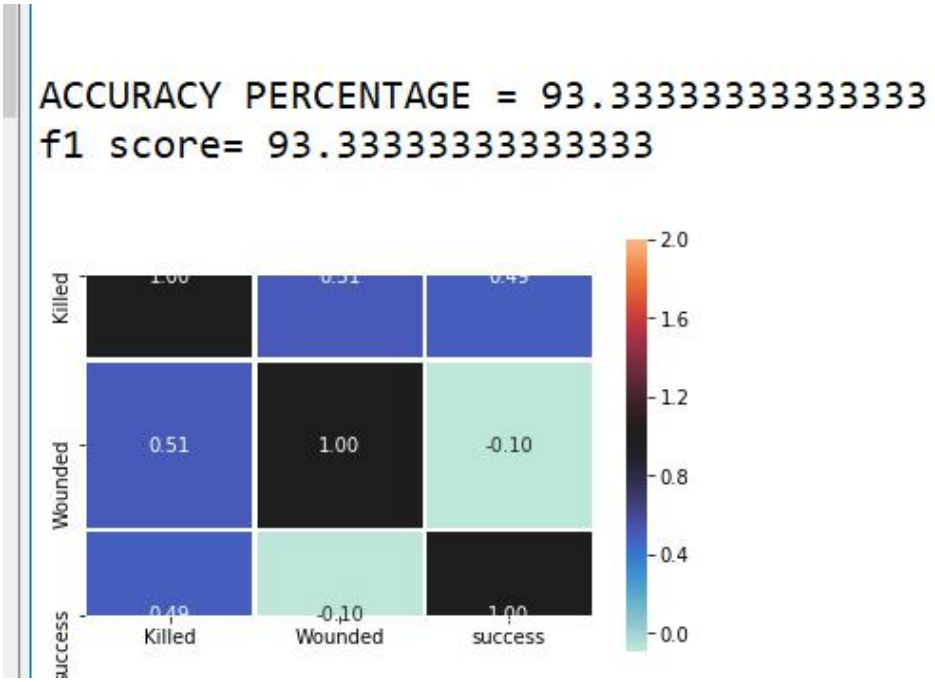
So with below graph one can say that "Probability to die by terrorist attack" is highest in Afghanistan.

Gaussian naive Bayes algorithm implementation

The Naive Bayes classifiers are working based on the Bayes' theorem, which describes the probability of an event, based on prior knowledge of conditions be related of conditions to the event. It is a very simple and fast classifier and works sometimes very good, and even without much effort you can get a okay accuracy.

```
df=df[df.AttackType == 'Assassination']
df=df[['Killed','Wounded','success']]
df.index = pd.RangeIndex(len(df.index))

features = df.drop(["success"], axis=1).values
target = df["success"].values


clf = GaussianNB()

X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0
######
clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)

a = accuracy_score(y_test, y_pred)
print("\n\nACCURACY PERCENTAGE =", a * 100)

# let's check the f1 score
b = f1_score(y_test, y_pred)
print("f1 score=", b * 100)

sns.heatmap(df.corr(), vmax=2.0, center=1, fmt='.2f',
            square=True, linewidths=1.50, annot=True)
plt.show()# -*- coding: utf-8 -*-
"""
Spyder Editor

This is a temporary script file.
```

Gaussian naive Bayes algorithm accuracy and F1 score with confusion matrix



Discussion: This screenshot shows results of Gaussian naive bayes algorithm result which is having 93.33% accuracy and 93.3% of F1 score.

NEWS REFERENCES

PolitiFact: Terrorism in the United States: key facts, patterns and trends (START in the News)
Los Angeles Times: I study terrorism. A white supremacist attack in my neighborhood woke me up to the danger at home (START in the News)
War on the Rocks: When Does Terrorism Have a Strategic Effect? (START in the News)

PR Newswire: 2019 Global Terrorism Index: Deaths From Terrorism Halved in the Last Four Years, but Number of Countries Affected by Terrorism is Growing (START in the News)
CNN: Report finds the Taliban were deadlier than ISIS in 2018 (START in the News)
Washington Post: The death of Baghdadi isn't the end of ISIS (START in the News)
CNN: Trump parrots talking points from Putin, Erdogan on Syria (START in the News)
The New York Times: Fact-Checking Trump on Syria, Erdogan and the Kurds (START in the News)

SELECTED PUBLICATIONS

Fact Sheet: Global Terrorism in 2018 (Fact Sheet)
Global Terrorism in 2018 (Background Report)
Terrorist Attacks against Religious Targets in the United States, 1970 – 2017 (Project Fact Sheet)
Terrorist Attacks Involving Package Bombs, 1970 — 2017 (Fact Sheet)
American Deaths in Terrorist Attacks, 1995 - 2017 (Project Fact Sheet)
Global Terrorism in 2017 (Background Report)
Heat Map: Global Terrorism Database, Terrorist Attacks 2017, Concentation and Intensity (Map)
American Deaths in Terrorist Attacks, 1995-2016 (Research Brief)
Ideological Motivations of Terrorism in the United States, 1970-2016 (Background Report)
Overview: Terrorism in 2016 (Background Report)
Heat Map: Global Terrorism Database, Terrorist Attacks 2016, Concentation and Intensity (Map)
Mass Casualty Explosives Attacks in Iraq and Afghanistan (Background Report)

Conclusion

Gaussian naive Bayes are used here for prediction because they are more robust to outliers, and as we have many outliers in spite of clipping the dataset, tree based models is the best bet we have! This is because, models like Logistic Regression or Neural Networks are very sensitive to outliers; consider a training scenario in a backpropagation network. Before backpropagating the error, all the error values are added together. The thing to notice is that error is averaged (divided by number of rows in dataset). So after few epochs the error due to rare classes turns out to be very low after averaging it. Due to which the network is unable to capture the error contributed by rare classes which makes them difficult to learn them. One way to learn such classes is that we can "penalize" the misclassified class, so that it will contribute more to overall error. Which is not the case with tree based models, beacause they literally don't look at how far is a data point. All they look is the "INFORMATION GAIN" (or decrease in RANDOMNESS of data) by splitting an attribute, which has more to do with how better it can classify target variable.

Moreover tree based models are fast learning algorithms and dataset given is also huge, it is advisable to use them. So, lets try tree based models: Random Forest, Decision Tree and Gradient Boosting Machines.

Here we have a multi-class classification problem. So f1 score is used as an evaluation metric with micro average. The dataset is highly skewed in terms of number of classes. So "macro average" will be a biased score as it is just average of individual Precision and Recall scores for each class, i.e., for classes with high frequency will have higher scores as compared to other classes which results in biased overall score.