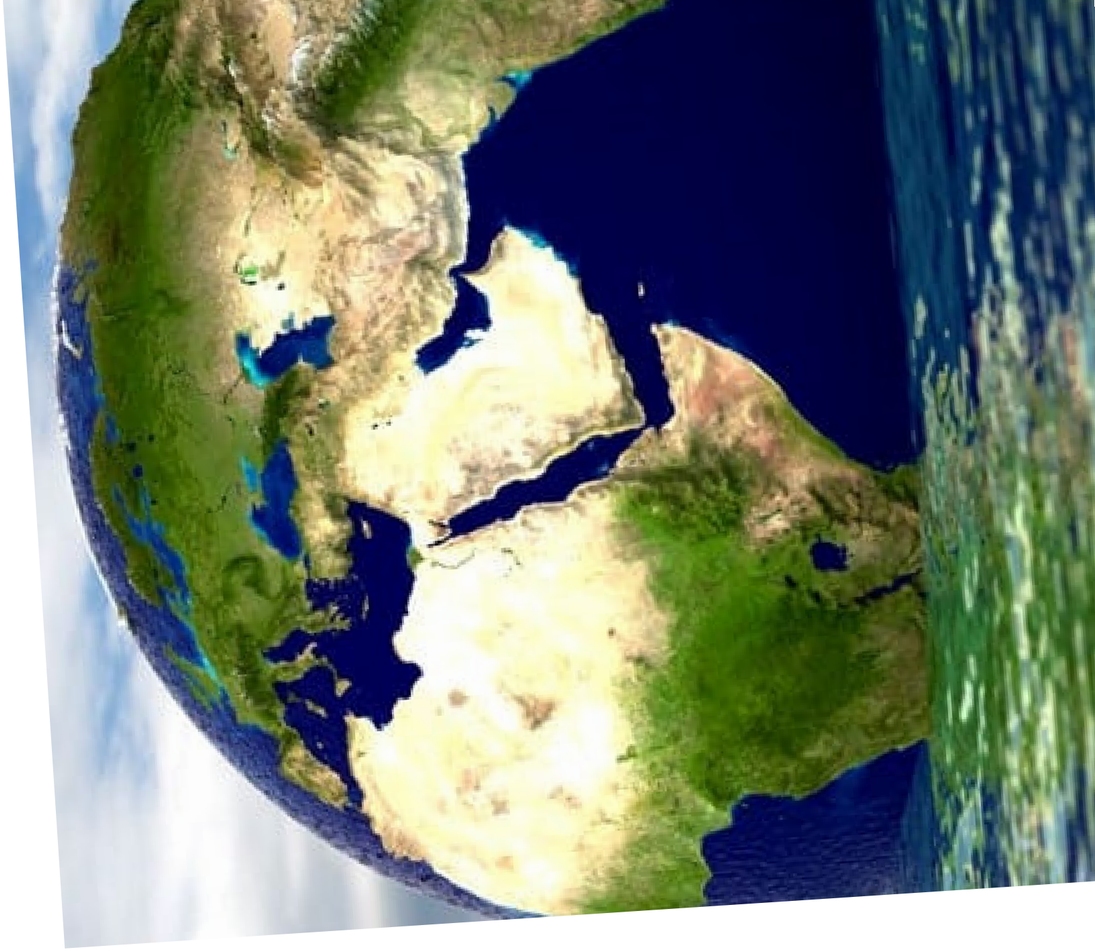
UNBDH Youth Track 2022

TEAM GASKIYA



Using Big Data and Data science to develop ideas and solutions to adress Global Challenges and help achieve Sustainable Development Goals

TEAM MEMBERS

- Vincent Theophane Meliga Naga
  Email: 699580621meliga@gmail.com
- Patience Akatuhwera
  Email: patience@aims.edu.gh
- Victor Osanyindoro
  Email: victor.osanyindoro@aims-cameroon.org
- Brenda Anague
  Email: brenda.anague@aims-cameroon.org
- Uriel Nguefack Yefou
  Email: uriel.nguefack@aims-cameroon.org

Outline

Introduction

Research Question

Data Used

Design an ML algorithm

How to improve the model

Conclusion

I) Introduction

- This report focuses on Global Annual Disaster Analysis from 1970 to 2021
- More and more, the world is exposed to natural disasters always growing
- 228 Countries were analysed, it was observed that a total of 8 billion people were affected, we had 4 billion US dollars of damages and 4 millions people lost their lives.
- We made the analysis of our data using two main tools: Power BI and Python

II) Research Question

THEME:
Using Big Data and Data science to develop ideas and solutions to adress Global Challenges and help achieve Sustainable Development Goals; notably to support policies caused by:

1

The disruption to Global Value Chains and Economic Globalization due to disasters, conflicts, restrictions, blockages

2

The impact of Climate Change on society as part of monitoring SDG 13

3

The rise of food and energy prices affecting vicious cycles of poverty, hunger, and inequalities

III) Data used

III-1) Presentation of the Data

The dataset we have used can be downloaded from this link: https://docs.google.com/spreadsheets/d/107SV-14k0GHzO0UrmwwwTmCj2XxwIXkkojZKe81xjN8/edit?usp=sharing

01

The disasters are divided into groups and subgroups

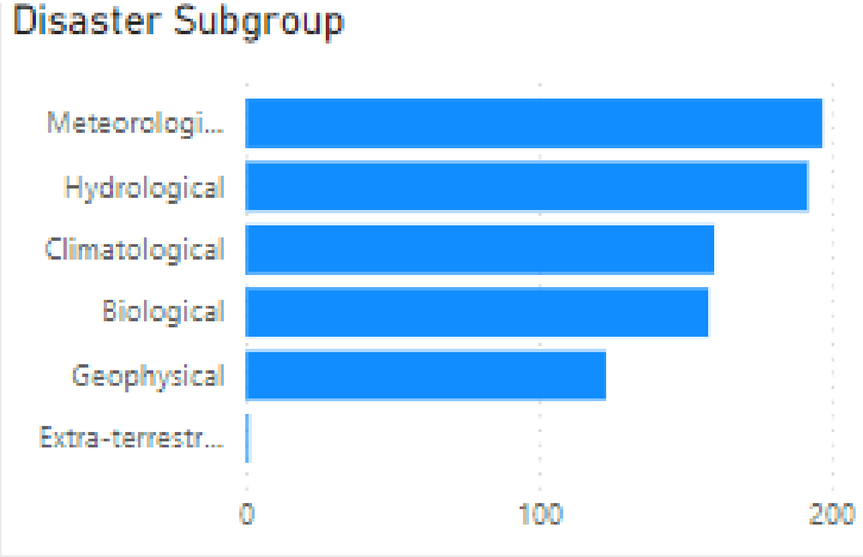
02

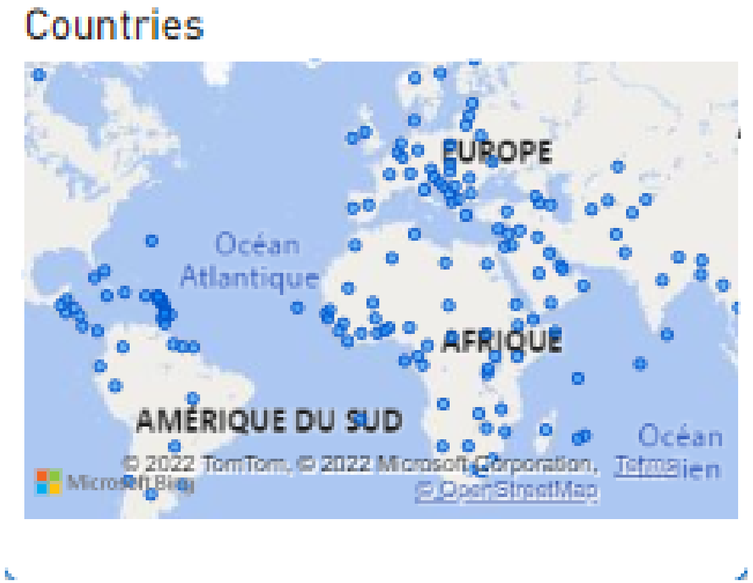
The disasters groups included meteorological, Hydrological, Climatological, Biological, Geophysical and Extra-terrestrial.

III) Data used

III-2) Data Visualization

For Quick visualization, the disaster subgroups are represented on the map of the globe and it is observed that the meteorological, Hydrological, and Climatical variables are dominating most countries.

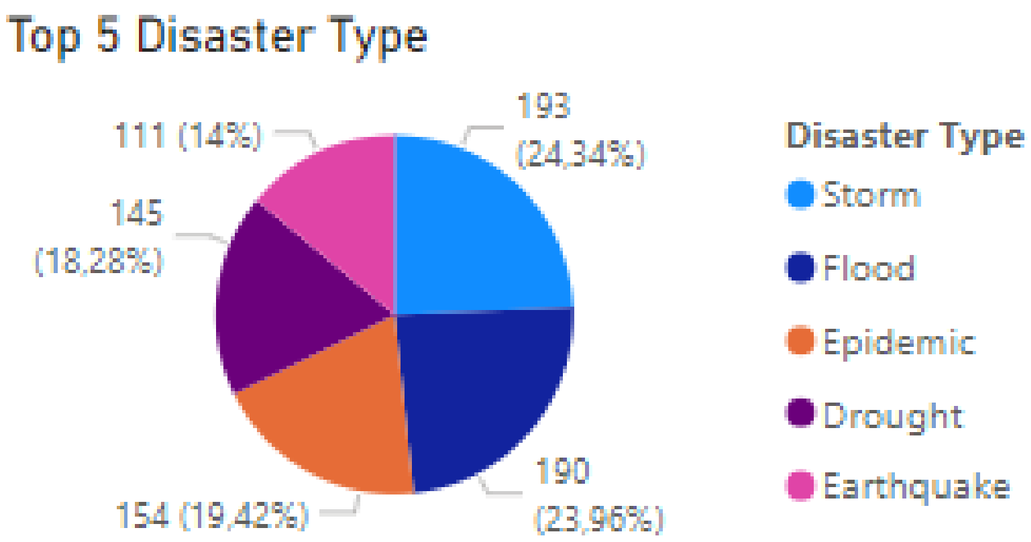




III) Data used

III-2) Data Visualization

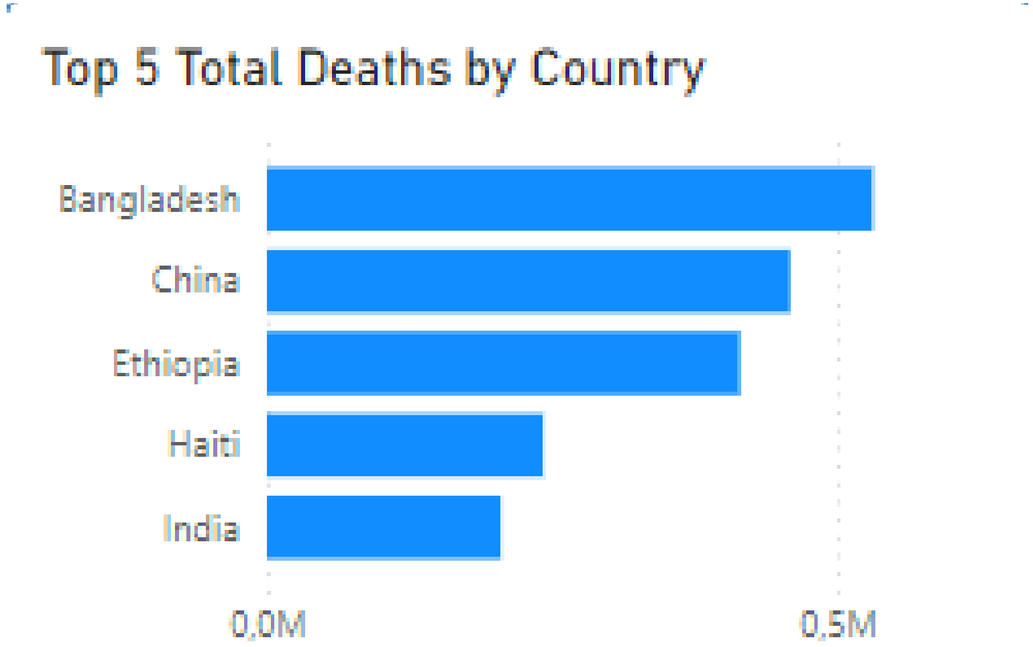
For the Top 5 disasters, we observe that most countries are affected by Storms(24.34%), then Floods(23.96%), followed by the Epidemic at 19.42%, Drought at 18.28% and Earthquake(14%)



III) Data used

III-2) Data Visualization

The Deathrates are recorded in the Top 5 countries that is Bangladesh, China, Ethiopia, Haiti, India respectively.



IV) Design an ML Algorithm

The goal of this part is to develop an algorithm that predicts the variable "OFDA Response" using Machine Learning. Our original column "OFDA Response" only had the value 'yes', so we replaced the blank space with 'no', now we can treated this problem as a binary classification problem.

We will apply several techniques to reach our goal:

- Data Cleaning
- Data Preprocessing
- Data Modeling

IV) Design an ML Algorithm

IV-1) Data Cleaning

In the Data Cleaning process, we have used the library Profile Report of pandas profiling to have the global summary of our dataset. From this Report, we have decided to drop the columns with a percentage of missing values >= 50%

- 'River Basin','Event Name','Origin'
- 'Associated Dis','Associated Dis2','Appeal','Declaration'
- 'Aid Contribution','Dis Mag Value','River Basin','No Injured','No Homeless'
- "Reconstruction Costs ('000 US$)","Insured Damages ('000 US$)"
- "Total Damages ('000 US$)","Country", "ISO" ,"Continent", "Location"

IV) Design an ML Algorithm

IV-2) Data Preprocessing

- Numerical columns: we have replaced the missing values by the median

- Categorical columns: we have converted categorical columns to numerical using the LabelEncoder

IV) Design an ML Algorithm

IV-3) Data Modeling

We applied several algorithms in order to find the best one after splitting the Data into 20%, 20%, 60% for the test set, Validation set and Train set respectively.

- Logistic Regression
- Support Vector Machines
- Random Forest Classifier
- Decision Tree Classifier
- Naive Bayes
- K Nearest Neighbor
- One vs Rest Classifier
- Bagging Classifier

IV) Design an ML Algorithm

IV-3) Data Modeling

From the table below, The Grid Search with Random Forest is the best algorithm with 86.9% of accuracy, followed by the One vs Rest Classifier with 86.49%

	Accuracy	Precision	Recall
AdaBoost Classifier	0.171277	1.00000	0.170394
Decision Trees	0.801064	0.43125	0.418182
K-Nearest Neighbor	0.807447	0.13750	0.338462
Grid search with Random Forest with Kpca	0.829787	0.00000	0.000000
Support Vector Machines	0.835106	0.05625	0.692308
Logistic Regression	0.836170	0.05000	0.800000
Naive Bayes	0.836170	0.08750	0.636364
Bagging Classifier with RepeatedStratifiedKFold	0.856383	0.35625	0.640449
Random Forest	0.864894	0.34375	0.714286
One vs Rest Classifier	0.864894	0.43125	0.657143
Grid search with Random Forest	0.869149	0.34375	0.753425

IV) How to imporove the model

From the last section, we have seen that the model was not perfect, there are some actions, we could perform in order to improve the model.

- Apply Feature engineering
- Test an ANN(Artifical Neural Network)
- Change the preprocessing technique
- Apply stacking technique

Conclusion

- 228 Countries were affected by Global natural disaster from 1970 to 2021
- Among the countries that have experienced natural disasters, the Philippines has experienced more than 300, followed by China, unlike Yugoslavia and Yemen which have suffered only one.

- Bangladesh is the country with more deaths(more than 500.000 deaths) and the Montenegro is the last one with just 1 death

THANK YOU
FOR LISTENING