Project Documentation

**Overview**

**Theme: Climate Change and Social Conflicts- to understand conflicts and societies and foster peace**

**AIInfreno &ModelMaker (dataset)**

* **Project Description**

The Emergency Events Database EM-DAT was created in 1988 as a joint initiative between the Centre for Research on the Epidemiology of Disasters (CRED) and the World Health Organization (WHO).

The CRED is now part of the Institute of Health and Society attached to the University of Louvain (UC Louvain). The EM-DAT database and project are primarily sponsored by the United States Agency for International Development (USAID).

EM-DAT contains data on the occurrence and impacts of over 26,000 mass disasters worldwide from 1900 to the present day. The database is compiled from various sources, including UN agencies, non-governmental organizations, reinsurance companies, research institutes, and press agencies. It is available in open access for non-commercial use and managed and distributed by the Centre for Research on the Epidemiology of Disasters (CRED) with the support of the United States Agency for International Development (USAID).

The main objective of the database is to serve the purposes of humanitarian action at national and international levels. The initiative aims to rationalize decision-making for disaster preparedness and disaster risk reduction strategies, as well as provide an objective base for vulnerability assessment and priority setting.

* **Problem statement**

The problem is to predict whether a state of emergency will be declared in an African country that experiences a specific type of natural disaster or not.

* **Data Source and Specification**

The dataset is gotten from the EM-DAT page - <https://public.emdat.be/data>

EM-DT focuses on major disasters. It globally records at the country level human and economic losses for disasters with at least one of the following criteria:

10 fatalities;

100 affected people;

a declaration of state of emergency;

a call for international assistance.

EM-DAT adopts a hierarchical classification. The database classifies disasters into two groups of hazards: natural and technological. The natural group is further classified up to four additional levels following the 2014 IRDR Peril Classification and Hazard Glossary. The technological group is less detailed and comprises three main types: transport, industrial, and miscellaneous accidents**.**

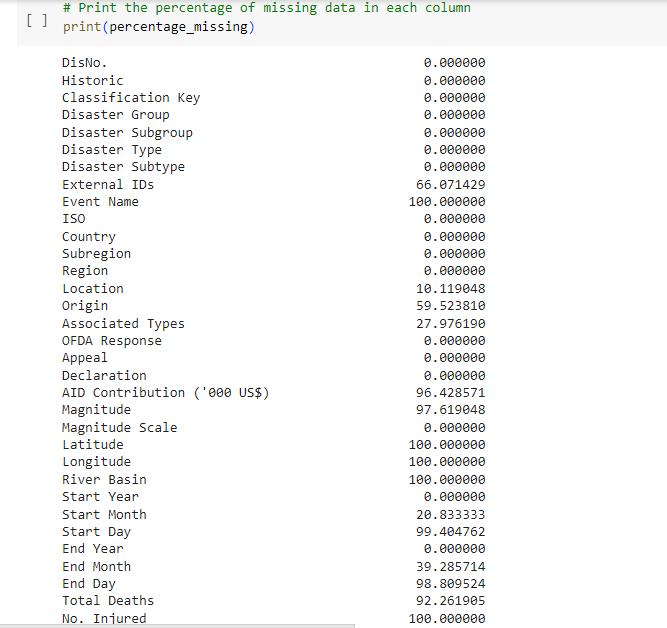
We are given a list of natural disasters in African countries between 2000 – 2023, the intensity, total number of people with physical injuries, trauma, or fatalities (dead and missing). The request for international assistance from the affected country, The total amount (in thousands of US$ at the time of the report) of contributions for immediate relief activities to the country in response to the disaster, costs of replacements of lost assets and the entry date of the events in EM-DAT records.

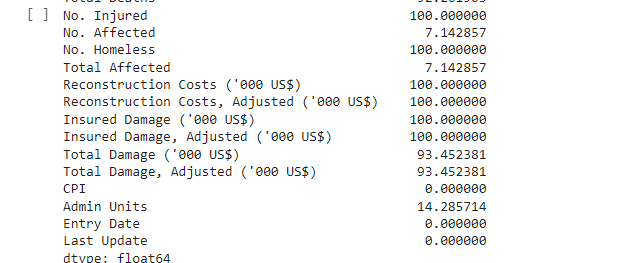
There are just two possible outcomes of the declaration which is either Yes or No.

* **Exploratory Data Analysis**

For the purpose of exploratory data analysis, the three subgroups were created and tasked with the responsibility of Pre-processing, Model creation and Model evaluation respectively.

**Missing Data Analysis**





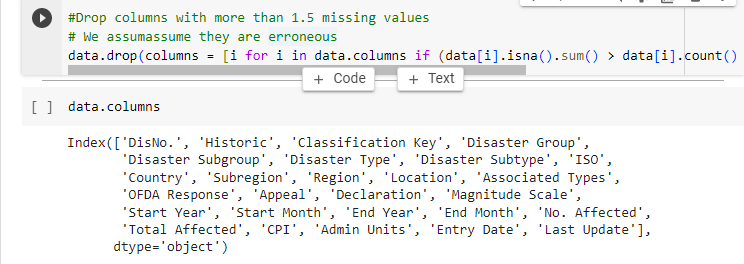
From the dataset, 66% of **External IDs,** 100% of **Event Name**, 10.12% of **Location**, 59.52% of the **Origin**, 27.98% of the **Associated Types**, 96.43% of the **AID Contribution**, 97.62% of the **Magnitude**, 100% of the **Latitude**, 100% of the **Longitude**, 20.83% of the **Start Month**, 99.40% of the **Start Day**, 39.23% of the **End Month**, 98.81% of the **End Day**, 92.26% of the **Total Deaths**, 100% of **No. Injured**, 7.14% of **No. Affected**, 100% of **No. Homeless**, 7.14% of **Total Affected**, 100% of **Reconstruction** **Costs** **(‘000 US$)**, 100% of **Insured Damage (‘000 US$**), 100% of **Reconstruction Costs**, **Adjusted** **(‘000 US$)**, 100% of **Insured Damage Adjusted (‘000 US$)**, 93.45% of **Total Damage** **(‘000 US$)**, 93.45% of **Total Damage, Adjusted (‘000 US$),** 14.29% of the **Admin Units** columns were missing.

**Data Cleaning**

On identification of the columns with missing values in the dataset, the next step is to remove the unfavorable variables for our model to perform better. The cleaning of the dataset was divided into steps.

***Step 1:***

Columns with more than 1.5 missing values were dropped from the dataset.



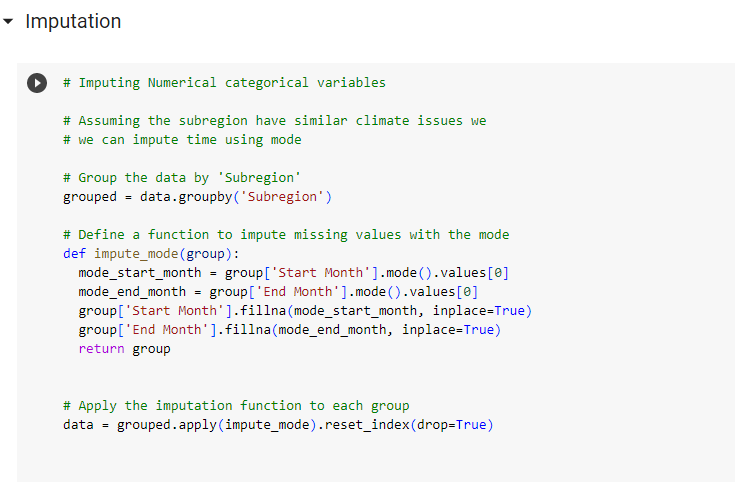
***Step 2:***

Some categorical variables were dropped from the dataset.



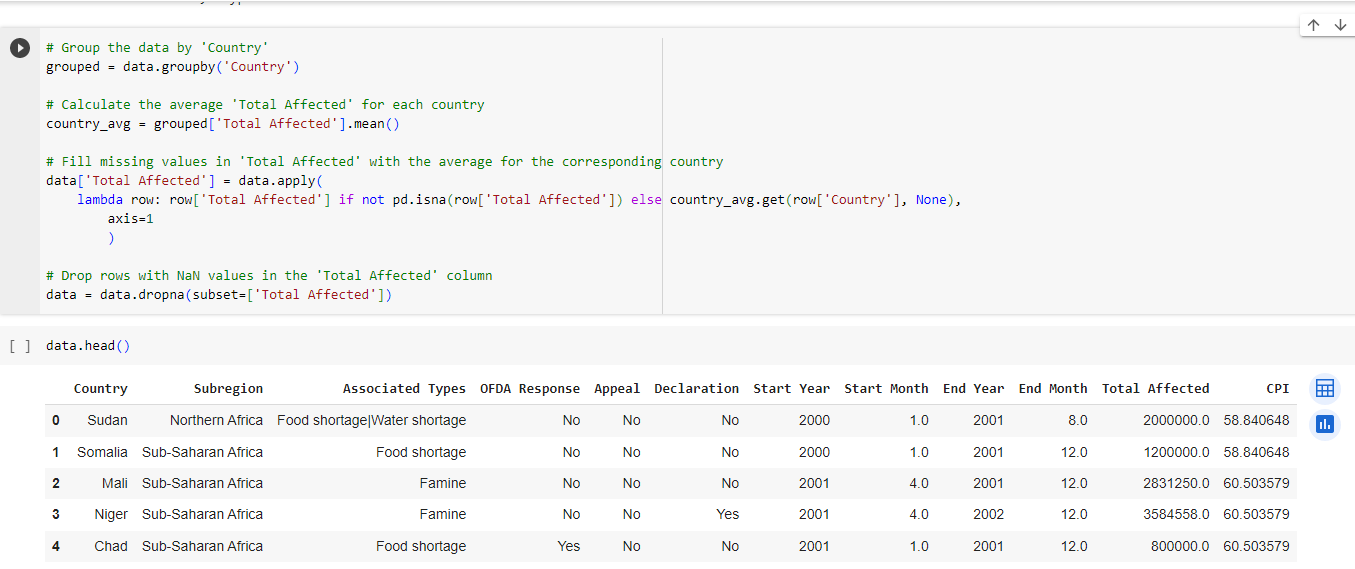
***Step 3;***

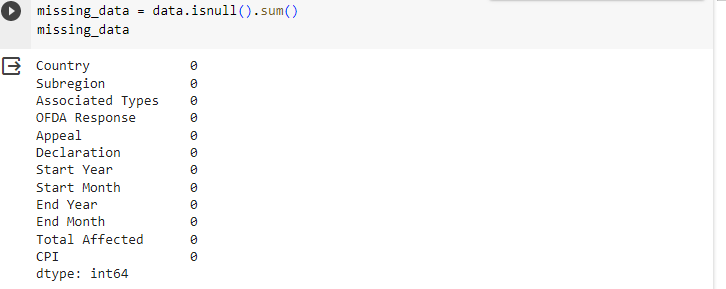
Grouping the dataset by subregion and filing the null values in the Start Month and End Month columns with the mode.



***Step 4;***

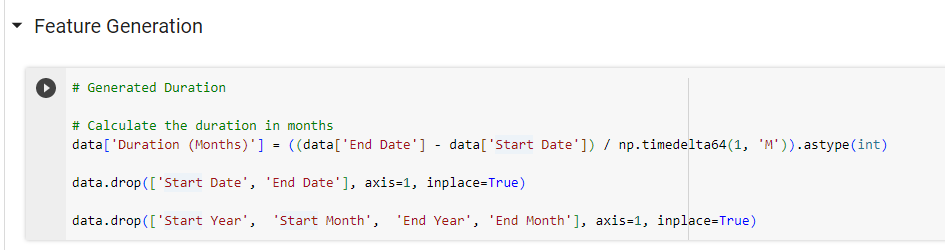
Group the dataset by country and calculate the average for the Total Affected column for each country and then fill the missing value with the average. After which the rows with missing values are dropped.

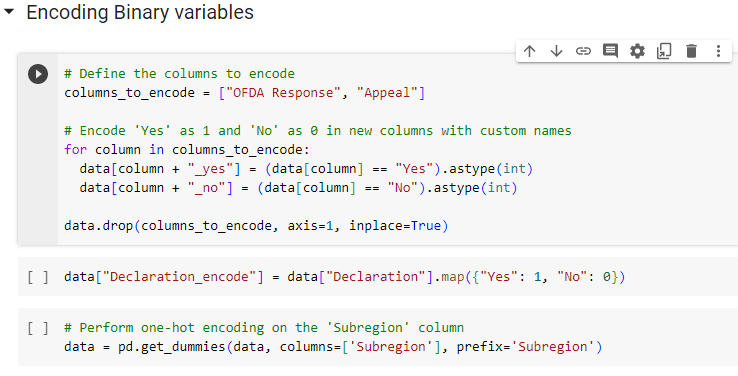




* **Encoding and Feature Engineering**

Encoding is the method of converting categorical values into Numerical equivalents. Alongside with Feature Engineering, it enables the design of artificial features in a dataset. These steps are very important in machine learning.

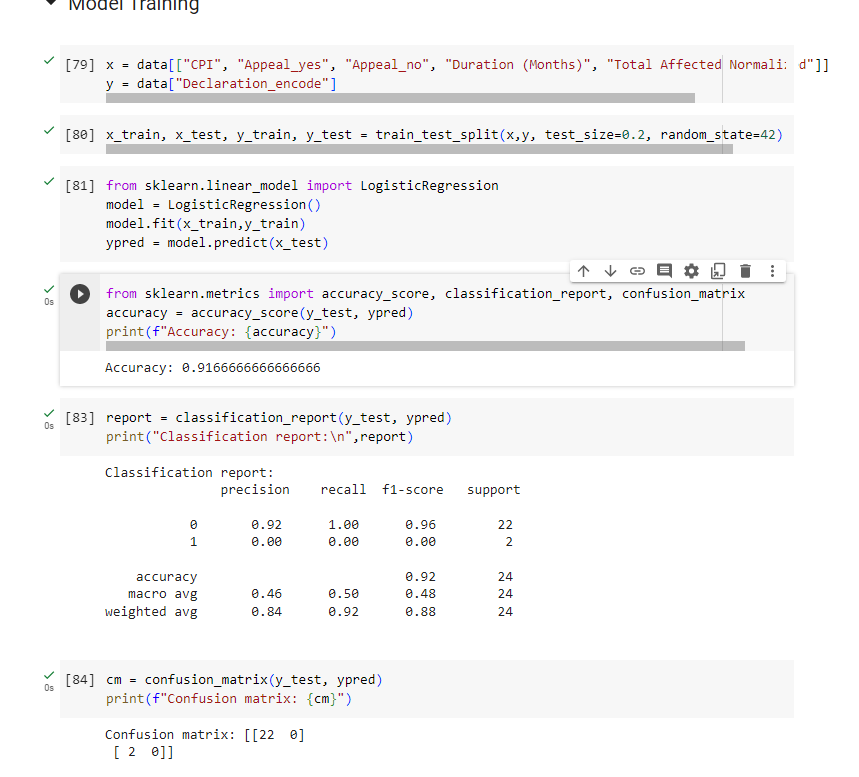




* **Building the Model:**

Now we have our dataset cleaned and ready for model implementation. We tested various models and datasets to get the best accuracy.

* **Building the Model:**



The baseline model was built using logistic regression. It showed an accuracy report of 91.6%

It also serves as our best model for dataset. It generated a precision of 0.92, F1-score of 0.96 and support of 22. A precision score of 0.92 depicts there are minimal false positives, and a recall score of 0.96 indicates the model is capturing most of the relevant cases. F1 -score is the harmonic mean of the level of precision and recall. A value of 0.98 showcases a strong balance between positive predictions and capturing all relevant instances. In summary, the model is accurate, has high precision, and effectively identifies a significant portion of the relevant instances in the dataset.

* **Conclusion:**

Using the model for predictions, we were able to make 22 correct predictions for Declaration of state of emergency in an African nation following a natural disaster.