**CLIMATE-INDUCED DISASTER PREDICTION**

**TEAM:-**

ADARSH KUMAR

ADARSH PATHAK

ANURAG TIWARI

**AIM and brief work TIMELINE:-**

The project aims to predict climate-induced disasters and floods using machine learning and deep learning models.

The work timeline can be divided into 3 parts.

* For this project, we read many papers, about general climate disaster predictions.
* Before finalizing the direction of work, we also worked on ArcGIS, tried to visualize the points, worked on raster data points, and worked on DEM models; as earlier, we tried to work on landslide prediction.
* Still, upon reading the papers, we concluded that it requires many parameters and is tough to build upon, so we, under the supervision of our mentor, decided to work on floods and chose Kerela as it has a lot of data points thus, predictions can be valid.

**PAPER DETAILS IN BRIEF:-**

The model in the paper aimed to predict floods in the Ontario region of Canada by using some climate parameters to train the model to be used.

In this paper, a deep learning (neural network) model for CID prediction is developed by linking historical disaster records to different climate change indices.

The paper is divided into two main parts,

1. The first part involves the general model structure that is generic enough to be employed to predict any class of CID in any location, given the availability of the influencing spatiotemporal climate data.
2. The second part demonstrates the applicability of the developed model using Ontario’s disaster rerecords and relevant climate change indices data.

This work is considered the first step in CID prediction, based on historical disaster data, global climate models, and climate change metrics, to maximize urban resilience and mitigate CID impacts on cities worldwide.

**IMPLEMENTATION:-**

So on similar lines, we attempted to obtain the similar parameters used in the paper and apply them to the Indian scenario in the state of Kerala. Since the dataset was maximum for Kerala, it was preferred. Although with sufficient training data, any region of India could be used.

Now there were many parameters used in the paper, but we decided to use the following set of parameters :-

Now the parameters used by us are listed:

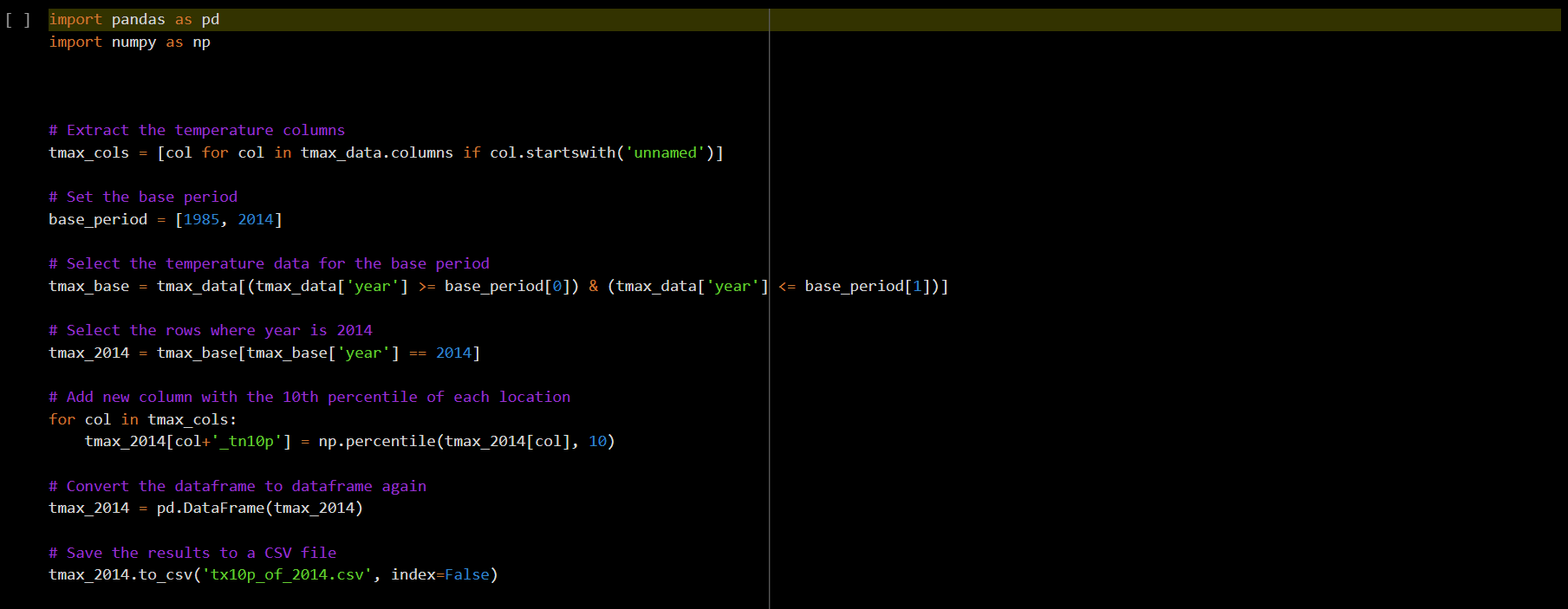
* DTR (Daily Temp Range)- this was the parameter which depicted the extremism of the weather and is relevant in the flood scenarios
* CWD-this parameter related to rainfall and thus was important to consider
* TN10p
* TX10p

Some parametrs like FROST DAYS were irrelavant for KERELA as its temperature never drops below 0 so no point of counting days below 0C.

1. **DTR** (daily temperature range) (Expert Team on Climate Change Detection and Indices 2009; Wazneh, Arain and Coulibaly 2017) represented by Eq. 1, where TXij and TNij are the daily maxima and minimum temperature, respectively, on day i in period j.

1. **CWD** (maximum length of the wet spell) (Expert Team on Climate Change Detection and Indices, 2009; Wazneh et al. 2017) which is calculated by counting the largest number of consecutive days where RRij≥1 mm, where RRij is the daily precipitation amount on day i in period j.
2. **TN10P** (Expert Team on Climate Change Detection and Indices 2009; Wazneh et al. 2017) refers to the percentage of days when TNij<TNin10, where TNij is the daily minimum temperature on the day i in period j and TNin10 is the calendar day 10th percentile centred on a 5-day window for the base period 1961–1990.
3. **TX10p**: Percentage of days when TX<10th percentile. Now, for calculating these parameters’ python was used and the codes will be attached below. Also, the use of excel was done.

**The way to calculate the value of different parameters is given in the readme file where an explanation of how it can be expanded for the different regions is also given.**



A sample code is provided for reference.

To incorporate all weather conditions, regions surrounding Kerala were taken, and some points of Rajasthan were included to incorporate extreme conditions.

This was done to remove the bias of the model towards Kerela.

The attempt was made to collect sufficient yes and no points for the flood to ensure data is not skewed. Also, in the Canada Model, some other parameters were used, but they were irrelevant to the Indian scenario, so they were dropped.

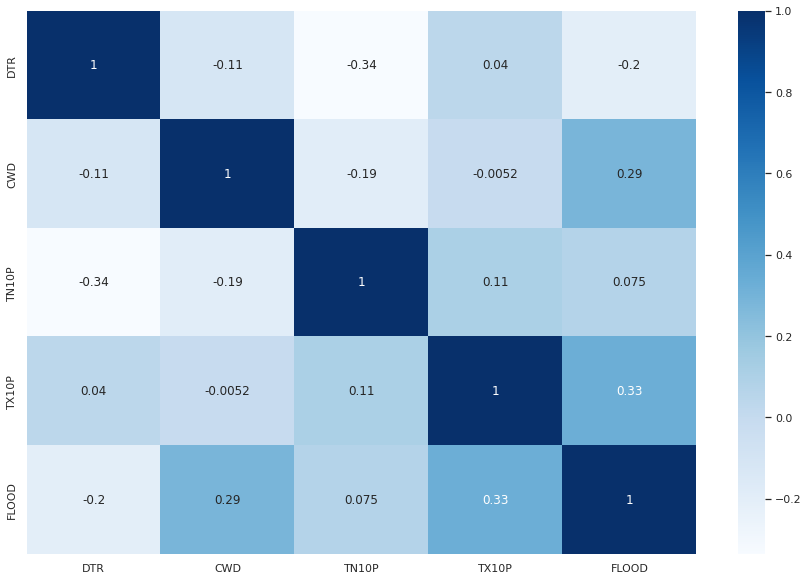
**RESULT: -**

The results and complete code are in this colab notebook.

<https://colab.research.google.com/drive/1m18GnPM47RqbfS9IS0oTHc9IeTayDABN?usp=sharing>

The snippets of the correlation of parameters with output and results of some models are provided below.

**CORRELATION VALUES: -**

We observed that TX10P and CWD give the maximum correlation with the output variable (FLOOD)

**P-VALUES: -**

In statistics, the p-value is the probability of obtaining a test statistic as extreme as the one observed, assuming the null hypothesis is true. The null hypothesis is typically a statement that no effect or relationship exists between two variables or groups being compared.

The p-value is used to determine the statistical significance of the results of a hypothesis test. If the p-value is very small (usually less than 0.05 or 0.01), it suggests that the observed result is unlikely to have occurred by chance, and the null hypothesis can be rejected. On the other hand, if the p-value is relatively large, it suggests that the observed result is not unusual, and the null hypothesis cannot be rejected.

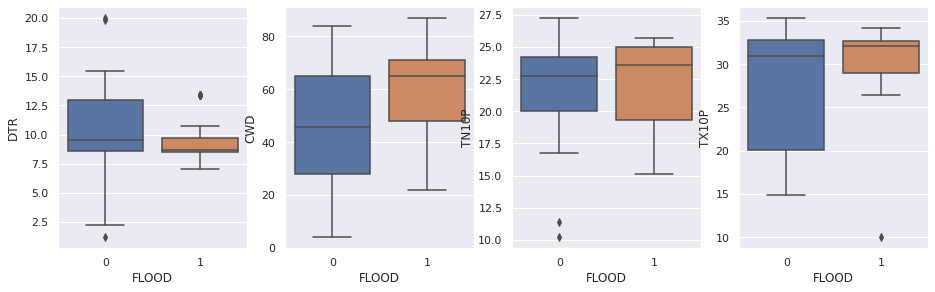
So our null hypothesis was that there was no discrepancy in the data at the time of it being collected and it was independent of all the parameters.

* DTR P VALUE: 0.00715
* CWD P VALUE: 0.00215
* TN10P P VALUE: 0.19678
* TX10P P VALUE: 0.01739

So we can see the only TN10P had a bit high p value but since it was not very high we will use it currently.

**BOX PLOTS AND OUTLIER ANALYSIS: -**

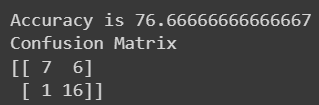
This shows if there are outlier or not



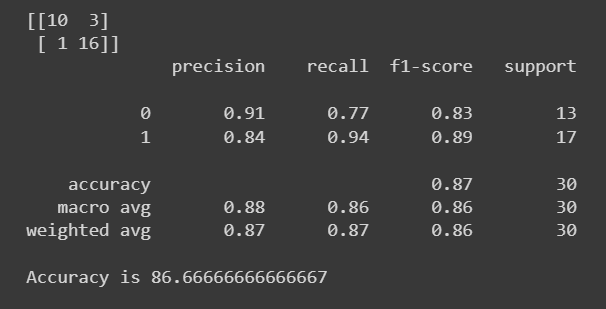
There were some outliers in DTR and TN10P but were less in quantity.

**# ACCURACY AND CLASSIFICATION REPORTS OF MODELS: -**

**LOGISTIC REGRESSION: -**



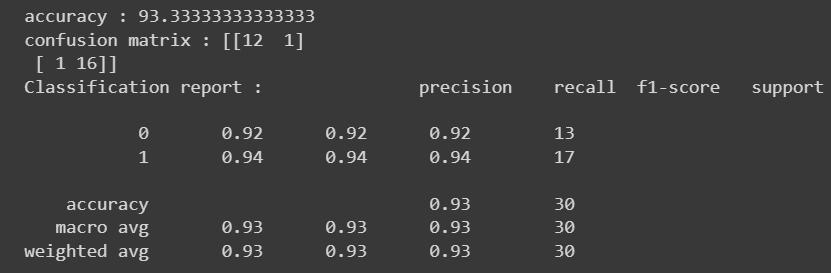
**DECISION TREE: -**



**XGBOOST (WITH UNSAMPLING): -**

Upsampling is a technique used in statistics and machine learning to address imbalanced datasets, by increasing the number of samples in the minority class.

Upsampling involves generating new samples for the minority class by replicating existing samples or creating synthetic samples based on the existing ones. This is often done when the dataset has a significant class imbalance, where one class has many fewer samples than the other.



**CONCLUSION**: -

* Our model did well and achieved high accuracy and the reason for not using NEURAL NETWORKS WAS THAT THERE WERE LESS DATA POINTS SO IT CAN OVERFIT EASILY AS COMPARED TO OTHER ML ALGOS.
* Our models and techniques to calculate and fill the data worked well, as seen from the p-values.
* Our choice of parameters also worked well, as seen from the correlation values.
* Also, from upsampling, the accuracy increased with XGBoost, so it also worked well.
* Thus, our work can be expanded for different regions also.