Christian Laurio
Theo Guilbert-Hattermann

REPORT

Project F7: FLOODING IN INDIA

I. Business understanding

1. *Identifying your business goals*

"There is no doubt that the flood situation this time is very serious and the frequency of rains is increasing significantly", *BBC: Assam: India floods destroy millions of homes and dreams (2022, https://www.bbc.com/news/world-asia-india-61862035*).

Floods have always been a relevant topic in India, especially due to the “Indian monsoon” phenomenon. This climatic event is a major wind system that seasonally reverses its direction. In India, the wind blows from the northeast during cooler months and reverses direction to blow from the southwest during the warmest months of the year. At the same time, in early summer, the lands and the Himalaya mountains heat up, so that hot air climbs up. Consequently, when the southwest wind meets this warm air, it creates large amounts of rainfall in the country during June and July.

Many factors contribute to flooding, but experts say climate change caused by global warming makes extreme rainfall more likely. India has experienced increasingly extreme weather in recent years, where unrelenting rains came just weeks after an extreme heat most of north India. For instance, in June 2022, 32 of its 35 districts have been affected by the tremendous floods, killing 45 people and displacing more than 4.7 million over one week.

This project aims to analyze and construe floods in India over the years, in order to identify the relevant features and then build a predictive flood model. Thanks to these information, we could extract some insights for policy makers and apply flood control measures at the end. The final predictive model must indicate if floods will occur or not, based on a bunch of parameters calculated throughout machine learning.

2. *Assessing your situation*

In order to do that, we will use three different datasets about floods and rainfalls in India:

- Monthly Rainfall Index and Flood Probability → https://www.kaggle.com/datasets/mukulthakur177/kerela-flood
- Flood Risk in India → https://www.kaggle.com/datasets/s3programmer/flood-risk-in-india
- Rainfall in India → https://www.kaggle.com/datasets/rajanand/rainfall-in-india?select=rainfall+in+india+1901-2015.csv

Each of them has interesting parameters which will help us to create a predictive model by combining them and using machine learning. This project started the 4th November and will end the 13th December. It may be possible that data are not fully complete, or that data from different datasets cannot be mixed (registered time or places are not the same for instance). This issue could delay our project and make the analysis trickier.

3. *Defining your data-mining goals*

The first thing we will need to do is to combine our three datasets in a single one. Afterwards, we will clean it by removing all unnecessary data. An important phase of the project will be then to identify, at first sight, which parameters are the most relevant for flood events. Next we will proceed to machine learning (random forests, linear regression, SVM…) and create some shrewd diagrams (data visualization), in order to emphasize the data and make them valuable. Finally, we could conclude to a reliable predictive model, where all parameters will have their own threshold. Then, the model will give the probability (in %) of the potential flood regarding the values of the parameters.

II. Data understanding

1. *Gathering data*

Three (3) datasets relating to the monthly rainfall index and other meteorological data from India were gathered from Kaggle. We searched Kaggle for datasets related to our objectives in predicting the probability of flooding in certain areas in India based on its historical monthly rainfall index. We were able to find and select three sources as stated in item 2 of the previous section. The datasets are open access and freely downloadable in csv file format.

2. *Describing data*

The datasets provided by Kaggle are already formatted in a tabular form in such a way that parsing or manual encoding are not needed. Here are the descriptions of each dataset we are considering in this project.
kerala.csv - this dataset came from the “Monthly Rainfall Index and Flood Probability” kaggle page published 4 years ago. The file contains the monthly records of rainfall index of Kerala from India from 1900-2018. The datafile itself is 10.3 kB and contains 16 columns namely SUBDIVISION, YEAR, one for each month, ANNUAL and label column whether it floods or not. The dataset has 118 entries or observations. The data type are mainly continuous values of rainfall index each month.

rainfall in india 1901-2015.csv - this dataset came from “Rainfall in India” Kaggle page. The page contains two files of data but only one will be used in this project. It also contains the monthly rainfall detail of 36 meteorological sub-divisions of India. It is an extension of the previous dataset of Kerala. The file is 528 kB and contains 4116 entries. The dataset has 19 columns but the last 4 columns will be disregarded as it is the quarterly average per year. The values are also numerical and continuous data.

flood_risk_dataset_india.csv - this dataset from "Flood Risk Prediction Dataset in India" in Kaggle is a synthetic dataset designed to aid in the development and evaluation of predictive models for flood risks across various regions of India. This dataset includes a diverse range of features that encompass meteorological (such as rainfall index, temperature, humidity) , geographical, hydrological, socio-economic, and historical flood data. This dataset is different from the two other previous as it also covers non-meteorological information relating to flooding. It contains 14 features and has 1000 entries. The file is 1.85 MB big.

3. *Exploring data*

The two dataset namely kerala.csv and rainfall in india 1901-2015.csv are very similar from one another. These two dataset can be merged into one file. Both contain the monthly rainfall index values of different months and years. The value ranges from 0 to 1098. Some entries in the rainfall in India 1901-2015.csv have NA values that need to be handled. There were no empty or missing values. Upon initial inspection of the data, some months have higher degree of rainfall index values and other months are lower. This could be attributed to the season during the year. The average yearly rainfall index are also given and these ranges from 1000 to 2000. This indicates a non normal distribution of the data and that needs to be looked for.

For the last dataset, this contains not only the rainfall index data but also other relevant features when it comes to predicting flood probability such as location, type

of soil and population. However, the dataset does not include the names of the subdivision but location coordinates are given. This gives us a problem in merging it with the previous two datasets. But this could be used in a data visualization model using a dashboard.

4. *Verifying data quality*

The number of entries are big enough to support the goals of the project. The combined dataset has at least 1000 entries. So far our main feature in training the model is the monthly rainfall index per year and per subdivision. We have one dataset that contains other features that might be relevant to predicting flooding probability however, our challenge is the merging of this dataset with other two with respect to the name of location. In order to check the quality of data, we will first conduct some preliminary data exploration such as checking for missing and erroneous values. When dealing with NA or missing data, we will see if it fits to perform data imputation when necessary. Data visualization will give us initial insights about the data. We will perform exploratory data analysis prior to developing and training the model. There might be a possibility to perform data pre-processing to standardized scale and normalize the values prior to the model training.

III. Planning

Tasks	Due Date	Description	Assigned To	Time Planned/Spent
Searching and retrieving datasets	11th November 2024	Searching datasets on Kaggle website which we could use to create our predictive model.	Théo, Christian	3h
Creation of the Github's page project	2nd December 2024	Create the Github's page for our project and invite instructors to the repository.	Christian	20m
Merging of all datasets	3rd December 2024	Combine the three datasets in a single	Christian	1h

		one.		
Cleaning data and exploratory data analysis	4th December 2024	Removing all unnecessary data and preparing them for the predictive model (machine learning).	Théo	1h
Training predictive model	6th December 2024	Build a predictive flood model thanks to machine learning techniques (random forests, linear regression, SVM…).	Théo, Christian	6h
Data Visualization	7th December 2024	Construe the results and show them across smart graphs for the poster.	Théo, Christian	2h
Poster creation	8th December 2024	Create the poster's template and then fill all the information and graphs.	Théo, Christian	5h
Suggesting solutions	8th December 2024	Suggest some ideas in order to help people to withstand floods and limit the damages.	Théo, Christian	2h
Poster submission	9th December 2024	Submission of the final poster.	Théo, Christian	30m