https://www.ncdc.noaa.gov/cdo-web/search
https://mausam.imd.gov.in/
https://www.imdpune.gov.in/lrfindex.php

Title: Predicting Precipitation with Machine Learning
Team: Jessica Ling, Kaitlyn Lane, Patrick Ledoit, Conor Luppnow

Task T: We propose to predict the precipitation in major meteorological subdivisions in India by month m as a function of historical weather information (monthly rainfall, average daily high/low temperature,).

Experience E: To do this, we will pull data from the Ministry of Earth Sciences, the India Meteorological Department, and potentially other data sources from 1901 to 2017. We will also compare our result to the predictions made by others discussed in 'Prior Work'

Performance metrics P: For ultimate evaluation of our contributions, we will use the mean absolute error of rainfall mm prediction. This is the same metric that is used by the two models cited in Prior Work, so we can easily compare our performance to theirs using mean absolute error. Since the existing models were to help predict flood or crop yield, the absolute rainfall mattered more than the percent error.

Prior work:

1. Deepthi Sudharsan, Isha Indhu S, Kavya S Kumar, et. al, "Rainfall Pattern Prediction". Github implementation at: https://github.com/DeepthiSudharsan/Rainfall-Pattern-Prediction-using-ML. This is a student project that predicts monthly rainfall in one meteorological subdivision in India (Tamil nadu) based solely on monthly rainfall from 1905-2015 using three methods: linear regression, lasso regression, and random forest. All models were imported from the sklearn toolbox.
2. V. Gaurav, "Rainfall Prediction". Github implementation at: https://github.com/vgaurav3011/Rainfall-Prediction This is a rainfall analysis project using monthly rainfall data from 1901 - 2015 (using the same dataset as above). The analysis is run on all provinces and briefly investigates linear regression (as well as with L1 and L2 regularization), random forest, SVR, and SGD. All models were imported from the sklearn toolbox.
3. Rajand Ilangovan, "Rainfall in India", Dataset at: https://www.kaggle.com/datasets/rajanand/rainfall-in-india This is the dataset that both repositories above use for their analysis. It has monthly, regional rainfall data from 1901-2015 across 36 meteorological sub-divisions in India.

Nature of main proposed contribution(s):

- The dataset used above has a large time scale (1905-2015) but is limited by both granularity (monthly) and type (mm rainfall only). We will use various other data sources to obtain higher-dimensional data.

    - We will possibly augment that dataset by verifying across another dataset and averaging the resulting monthly mm rainfall.
    - We will also add additional pertinent data, such as monthly high/low temperatures.
- They used linear regression, lasso regression, and random forest machine learning methods. We will use similar methods, changing parameters.
    - Generally, the projects above did an 80:20 split of training to testing data. We will use k-fold cross-validation.
- We will initially perform some exploratory data analysis and visually describe our data.

Why we care: Knowing the precipitation helps you prepare in the morning so you do not get rained on. Currently, weather predictions only predict ~10 days ahead, our model will predict farther into the future and help individuals plan around the weather. For example, it could help inform farmers of how much water they need to use to water their crops.

Which parts of the curriculum from this class do you expect to apply?:
We expect to use the machine learning methods from the earlier portion of the class, such as linear regression, lasso regression, and possibly a random forest regressor. We will also need to perform some data cleaning and data wrangling as the datasets we are using are not perfect. We will also use k-fold cross-validation in order to obtain good train/test/validation splits. As it stands so far, we do not intend to use any neural networks, but if there is a good enough reason to do so we may try using one.

Compute Requirements: We will use the Google Colab instances as well as personal machines in order to run our programs.

Expected challenges and risk mitigation: One challenge we will have to deal with is potentially a lack of data. Initial research leads us to believe that we can find some additional measures such as temperature for some areas of India. However, finding data for all the subdivisions of India that are formatted in the same way as the existing dataset may be difficult. We may have to manually aggregate weather data as several datasets give daily weather information over time.

Another anticipated challenge is handling unexpected external factors in weather predictions, particularly in considering how a predictive weather model may be different from 1905 to 2015. For one, data availability is a concern. We may want to limit the time-scope of our model, particularly since finding the data to augment current data (adding temperature, for example) is exponentially harder for years farther in the past. For another, weather patterns may have significantly shifted during that century, which may lead to certain systemic difficulties in creating a model that predicts accurately for all years within that century. This is an area that will require further investigation to clarify.

Ethical considerations and broader social impact: India and other areas are at high risk of floods during different seasons, additionally, rainfall is key to agriculture. Knowing periods of drought or high rainfall is very important for farmers to determine the right crops to grow, the best time to harvest, and precautions to undertake on the field. Thus, predicting rainfall can help communities predict and prepare for floods and plan their agriculture ventures.

(exempted from 2-pg limit) Work Plan over the next ~5 weeks:

PERSON (S)	TASK (S)	Wk0				Wk1				Wk2				Wk3				Wk4				Wk5			
		March									APR														
		S 17	M 18	W 20	Th21	S 24	M 25	W 27	Th28	S 31	M 1	W 3	Th4	S 7	M 8	W 10	Th11	S 14	M 15	W 17	Th18	S 21	M 22	W 24	Th25
All	Project Proposal			■																					
All	Data Set Search						■	■																	
All	Data Preprocessing, Merge, and EDA										■	■	■												
All	Feature Engineering														■	■									
All	Run Existing Models on New Data																		■	■					
All	Project Check In																						■		

Each task over the next 5 weeks will involve everyone. Each task will be broken down further and delegated.

(Supplementary materials if any, also exempted from 2pg limit, but not guaranteed to be considered during evaluation)