Project Proposal

Project Title: Application of Deep learning in annual temperature and rainfall prediction of India

Domain: Weather forecasting is the task of predicting the state of the atmosphere at a future time and a specified location. Traditionally, this has been done through physical simulations in which the atmosphere is modeled as a fluid. The present state of the atmosphere is sampled, and the future state is computed by numerically solving the equations of fluid dynamics and thermodynamics

Problem statement: Although the current weather prediction models used by metrological department produce good results, this physical model is prone to error due to noise and uncertainties in the dataset of the atmospheric conditions, and an incomplete understanding of complex atmospheric processes restrict the extent of accuracy in weather forecasting for a longer period of time which makes weather forecasts significantly unreliable.

Machine learning, on the contrary, is relatively robust to perturbations and doesn't require a complete understanding of the physical processes that govern the atmosphere. Therefore, machine learning may represent viable alternative to physical models in weather forecasting.

This is a regression problem involving time series analysis. My dataset has two major variables- temperature and rainfall. And they are divided into monthly minimum, maximum and mean. Both these variables are interdependent. Current month temperature would depend on temperature and rainfall of previous months and vice versa for current month rainfall. Hence, these would be regarded as continuous variables.

Solution: Temperature of an area is one of the factors that are directly proportional to total rainfall in an area. Similarly, amount of rainfall also decides temperature range in the area. Hence study of temperature (min, max and mean) and amount of rainfall for past years can help us to predict future temperature and rainfall.
Time series analysis will be required to solve this kind of problem. Many methods can be applied. Naïve approaches like Single Exponential smoothing, Holt's linear trend method can be used to initial study and then more advanced methods like Holt's Winter seasonal method, ARIMA, multivariate time series prediction using LSTMs RNN can also be applied if required.

Datasets and Inputs: Annual temperature and rainfall datasets can be accessed from data.gov.in. Annual mean, annual minimum, annual maximum, monthly average temperature and Jun-Sept total rainfall data are available on data.gov.in as linked below:
Jun-Sept rainfall in central india data
Annual Minimum/Maximum temperature data
Monthly average temperature

The minimum temperature is in the range of 16-20 and the maximum in the range of 28-32. Hence, they seem to be an average minimum/maximum of all the recordings across the region.

Dataset shape = 116 x 19.

Note: In India most of the rainfall occurs in months of June to September. Hence, data for total rainfall June-Sept would be approximated as total annual rainfall.

Benchmark model: Since very limited work has been done on application of machine learning in weather forecasting, it is very difficult of find a freely available well accepted solution. Hence, a a naïve simple study of 'Single Exponential Smoothing' on my dataset will be used as a benchmark model. This would give a base line score of my dataset and then dataset would be studied on more advanced methods as mentioned under solution heading.

Evaluation Metrics: Root mean squared error (RMSE) would be used to quantify error of my model. Also, a graphical comparison would also be used to view the differences between predictions and test dataset.

Project design:

1. Data-preprocessing – Data is already clean without any missing values. So, none of the rows will be dropped. In case of LSTMs, they are sensitive to the scale of the input data, specifically when the sigmoid (default) or tanh activation functions are used. Rescaling the data to the range of 0-to-1, (normalizing) would be required.
2. Train and test data – Since there are total 116 datasets available, 100 would be used for training and 16 for testing. With time series data, the sequence of values is important and I will split the ordered dataset into train and test datasets, first 100(year 1901 to 2000) as train and last 16(year 2001 to 2016) as test dataset.
3. Training the data – As mentioned under solution heading, model training would start from Single Exponential smoothing, Holt's linear trend method and more advanced methods like Holt's Winter seasonal method, ARIMA can be applied and then if required multivariate time series prediction using LSTMs RNN can also be applied If results fail to be reasonable with initial training methods.
4. Prediction: Annual minimum, maximum, mean temperature, monthly average and total annual rainfall.
5. Testing: RSME will be used to quantify losses when testing our trained model in test dataset.