Forecasting Monthly Rainfall with Classical Machine Learning: A Comprehensive Approach for Subdivision-Level Predictions

Aditi Sharma, Ayana, Bharti Bansal, Komal

Abstract

Accurate monthly rainfall prediction is crucial for various sectors, including agriculture, water resource management, and disaster preparedness. Stakeholders, particularly farmers, rely on precise forecasts to make informed decisions regarding crop planting, irrigation scheduling, and water conservation. Here, we have analyzed rainfall patterns and leveraged the results from there to decide how much data to use for training while predicting for a specific year. We aim to use different models for this and compare their performances. For more details, you can check the project repository at: GitHub Repository: ML_Project-24

Introduction

0.1. Context

Accurate rainfall predictions are crucial for various sectors, including agriculture, water resource management, and disaster preparedness. Stakeholders, especially farmers, rely on these predictions to make informed decisions regarding planting, irrigation, and resource allocation.

0.2. Significance

By improving the precision of rainfall forecasts, we aim to empower stakeholders with the insights needed to make better-informed decisions. Enhanced forecasting capabilities will support optimized water resource management and better preparation for rain-related challenges, ultimately contributing to improved agricultural productivity and resilience against climate variability.

Literature Survey

Prediction of Rainfall using Data Mining technique Over Assam

There are mainly two approaches to predict rainfall, Empirical methods and Dynamic methods. This paper implements Empirical statistical technique using Multiple Linear Regression (MLR). The dataset used here consists of local meteorological data from stations over Assam over 2007-2012. Finally the model used Max Temp., Min. Temp., Wind Speed and Mean Sea Level as predictors. They found 63% accuracy in variation of rainfall using MLR. Paper discussed that due to constraints on data collection wind direction is not used as predictor which could give more accurate results.

Predicting long-term monthly rainfall with an ensemble of climatic and meteorological predictors

The research paper compares ARIMA and five state-of-the-art ML models across 25 stations in East China. The study compares the performance of various ML models, including Random Forest (RF), against the traditional ARIMA model. The results indicate that ML models, particularly RF, outperformed ARIMA in predicting rainfall, although all models faced challenges with accurately forecasting extreme rainfall events. Key predictors identified in the study were local meteorological variables such as humidity and sunshine duration, along with a significant 4-month lag between the Western North Pacific Monsoon and rainfall. The study also highlights the scalability of the multi-ML ensemble framework, which can be applied to other regions to enhance forecasting accuracy. Future research could focus on integrating more climatic variables and improving the interpretability of ML models.

1. Dataset Details

1.1. Dataset Overview

We used 45,292 entries from the IMDAA reanalysis dataset (Rani et al., 2021) for weather forecasting over India, focusing on data from 1990 to 2020 from BharatBench. The original dataset contains over 4 crore entries, offering detailed meteorological information with a 0.12° spatial grid (meaning coordinates are taken at a gap 0.12°) and 24 vertical levels (latitudes) from 1979 onwards. This reanalysis dataset combines simulation results with observations, providing a continuous, high-resolution representation of atmospheric conditions, crucial for understanding weather patterns.

1.2. Final dataset

The final features are: updated_lat, updated_lon, year, month, APCP_sfc, TMP_2m, wind_speed, MTERH_sfc, wind_dir, HGT_prl, TMP_prl. Other than this we combined also combined some more features like Darwin and Tahiti SLP, SOI, Central Pacific OLR Index, NINO 3.4 SST. Combining all these features we finally made our dataset by grouping it monthly.

2. Data Preprocessing Techniques

2.1. Handling Missing Values

2.1.1 Identifying Missing Entries

Conducted a thorough analysis of the dataset to identify missing entries in dataset. Then, applied filling techniques to estimate these missing values based on existing data. major missing data was of longitude and latitude.

Used a systematic approach to generate plausible latitude and longitude values:

- **Latitude Range:** 5.04° N to 38.52° N
- **Longitude Range:** 65.04° E to 98.51° E
- For missing values, incremented existing latitude and longitude by 1.08 degrees, wrapping around within the defined geographical limits when the maximum value was reached. This approach ensured that all missing entries were filled logically, maintaining geographical relevance.

2.2. Conversions

- Time in initial 4 files was in UNIX encoding, we used pandas to convert it into seperate columns: 'year', 'month', 'day', 'day_of_week', 'week_of_year'.
- Used columns 'UGRD_10m' and 'VGRD_10m' to calculate 'wind_speed' and 'wind_dir'.
- Added column 'MTERH_sfc' to final dataset.

2.3. Combining old and new datasets

Our old dataset was not monthly grouped, so we converted it so that we now have data for each month of each year for each subdivision. The new datasets were calculated monthly. So, this monthly grouping of the earlier datasets was necessary. Final correlation matrix came out as:

2.4. Data Splitting

Divided the dataset into training and test sets to ensure robust model evaluation and prevent overfitting:

- Typical split ratios: 80% for training, 20% for testing
- For testing base model we used year 2020 data for testing and 1990 to 2019 data for training.

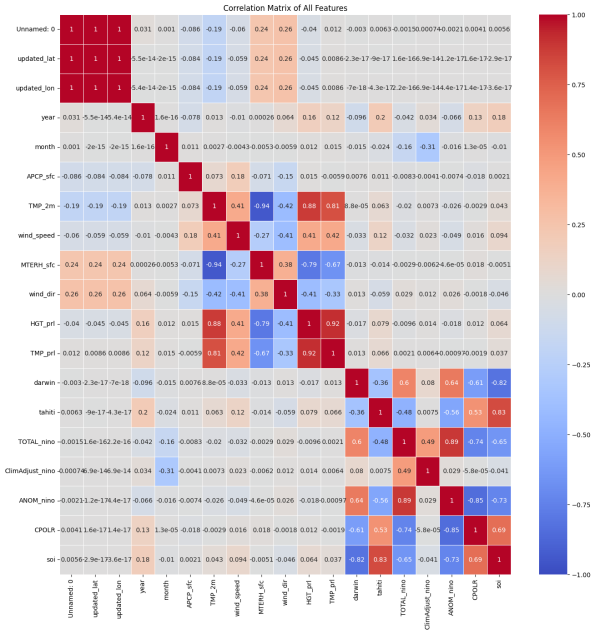

Figure 1. Correlation matrix of final dataset

3. Methodology and Model Details

3.1. Methodology

Ways of Rainfall Prediction

- Synoptic Weather Prediction: It refers to the observation of various objects during a particular period of observation. It is a traditional approach. The Meteorological Department produces a chain of synoptic charts to record the track of changing weather patterns, which are the basis for weather forecasting. It includes a vast collection of observed data obtained from most weather stations.
- Statistical Weather Prediction: This type of prediction classification is employed at the side of the numerical strategies. It makes use of historical climate data, assuming that the longer term weather will be similar to previous weather conditions. The main goal of this prediction is to determine which components of climate gives best prediction result for future prediction. As a result, the overall situation of the climate can be anticipated in this manner.
- Numerical Weather Prediction: The computer's power predicts the weather. Complex computer programs are used on supercomputers and estimate numerous climate parameters. The process of the equations used is not more accurate, and the weather's first stage is often partially known. Hence, weather forecasting will not be precise entirely

- Further There are two modelling approaches to predict rainfall - Empirical and Dynamical Methods. The empirical method is based on looking at actual weather data and how it relates to a variety of atmospheric variables. Regularly practical approaches are used to predict the climate, that is, ANN, regression, group method, and mathematical logic for data handling.
- The dynamical approach, predictions are generated by physical models based on system of equations that predict the future rainfall. Meteorologists have developed atmospheric models that approximate changes in temperature, pressure, etc. using mathematical equations.
- From the above prediction and modeling approaches we will use the empirical statistical method since we have historical weather data from 1990 - 2020 and its relationship to a variety of atmospheric variables.
- The analysis of rainfall trends over the years revealed significant variations approximately every 5 years, as observed through graphs and plots comparing actual rainfall data. Aggregating 3 years of data to predict the subsequent 4 years resulted in poor prediction performance, highlighting the challenges posed by changing trends. To address this, we utilized a correlation heatmap for feature selection, identifying the most influential features for rainfall prediction. These selected features were then used to train various models, including Random Forest, Decision Tree, Linear Regression, and MLP Regressor, to optimize predictive accuracy.

3.2. Model Details

We performed grid search to get the best parameters for each model and used those for our predictions.

- **Random Forest:**
  - Best Parameters: {'max_depth': 5, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 100, 'random_state': 42}
  - MSE: 43.62
  - RMSE: 6.60
  - R^2: 0.29
- **Decision Tree:**
  - Best Parameters: {'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 5, 'random_state': 42}
  - MSE: 61.25
  - RMSE: 7.83
  - R^2: 0.00
- **MLP:**
  - Best Parameters: {'activation': 'tanh', 'hidden_layer_sizes': (50,), 'max_iter': 1000, 'random_state': 42, 'solver': 'sgd'}
  - MSE: 61.75
  - RMSE: 7.86
  - R^2: -0.01

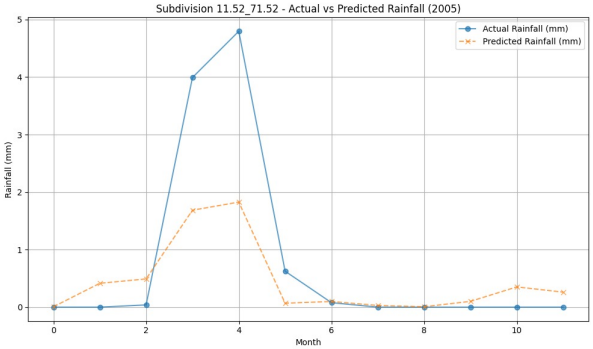

Figure 2. Using only 2 features

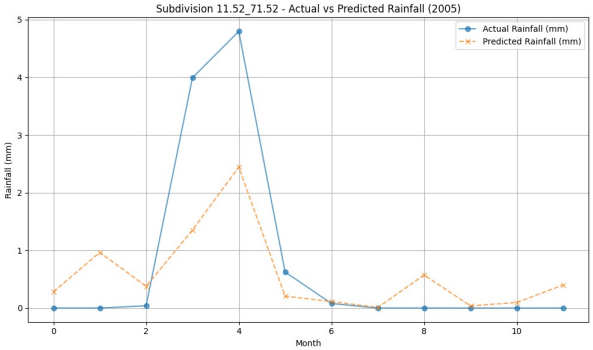

Figure 3. Using all features

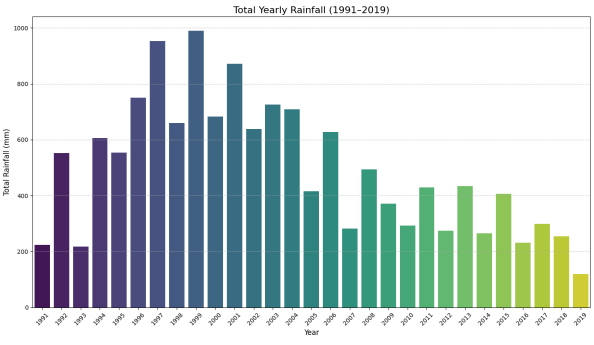

Figure 4. Yearly rain comparision

4. Results and analysis

Yearly rain comparison as seen shows the fall of rainfall from 1990s to 2000s and then to 2010s. This gives us the clear idea that using data from 90s to predict the rainfall for 2020 is a bad idea and will lead to bad predictions. This change in rainfall pattern across decades can be reasoned by looking at the adverse climate change the earth has been through with industrialization and urbanization.

Table 1. Average Metrics Across All Subdivisions (1990–1995)

Model	**Avg MSE**	**Avg RMSE**	**Avg R^2**
Random Forest	19.75	2.85	0.41
Decision Tree	26.57	3.29	**0.51**
Linear Regression	20.35	3.06	0.22
MLP Regressor	22.66	3.29	0.02
Best Model	**Decision Tree (Based on Avg R^2)**		

Table 2. Average Metrics Across All Subdivisions (1995–1999)

Model	**Avg MSE**	**Avg RMSE**	**Avg R^2**
Random Forest	45.97	4.84	0.25
Decision Tree	60.75	5.57	**0.27**
Linear Regression	45.35	4.55	0.19
MLP Regressor	55.71	4.62	`-inf`
Best Model	**Decision Tree (Based on Avg R^2)**		

Table 3. Average Metrics Across All Subdivisions (2000–2010)

Model	**Avg MSE**	**Avg RMSE**	**Avg R^2**
Random Forest	12.32	2.75	0.23
Decision Tree	22.33	3.39	**0.37**
Linear Regression	10.60	2.55	0.08
MLP Regressor	9.42	2.35	`-inf`
Best Model	**Decision Tree (Based on Avg R^2)**		

Table 4. Average Metrics Across All Subdivisions (2010–2020)

Model	**Avg MSE**	**Avg RMSE**	**Avg R^2**
Random Forest	3.50	1.54	0.20
Decision Tree	5.11	1.79	**0.27**
Linear Regression	2.05	1.20	0.13
MLP Regressor	2.36	1.22	0.00
Best Model	**Decision Tree (Based on Avg R^2)**		

- As evident from the figures, performing the Random Forest model using the two best features versus all features shows minimal difference in the R^2 value. This indicates that features other than wind speed and *tmp_2m* have lower correlations with the target variable and contribute less to the model's predictive performance.

5. Conclusion

5.1. Learnings

- We gained valuable insights into data analyzing and modeling for time-dependent datasets, improving predictive accuracy through feature enrichment from multiple datasets.
- By working with 31 years of rainfall data, we effectively managed large datasets by grouping, and identifying key patterns.
- Visualized unpredictable rainfall trends across time, locations, and types, enabling better comparisons and insights.
- Overall, the experience equipped us with a data-driven mindset and analytical skills to address similar problems in the future.

5.2. Key Findings

- The rainfall trends change significantly over five years, impacting prediction accuracy across different subdivisions. Using three years of data (e.g., 1995–1998 to predict 1999) provided the most accurate predictions compared to using data from four, three, or all 31 years, highlighting the importance of selecting an optimal time window for forecasting.
- One more trend observed was the alternating nature of rainfall; the months that received high rainfall in one year had the opposite trend and didn't get as much rain.

5.3. Challenges

- Faced problems due to missing rainfall data along with integrating multiple datasets while ensuring its compatibility.
- Model selection for time-series forecasting and tuning for optimizing performance was another challenge that we faced.

5.4. Contribution

- All members contributed equally to final model training, result analysis, presentation creation, and report preparation.

References

[1] A. K. Singh, A. Kumar, B. Ahuja, and K. Gupta, BharatBench: Dataset for data-driven weather forecasting over India

[2] J. Zhou, B. He, and X. Guo, A comparative study of extensive machine learning models for predicting long-term hydrological processes

[3] A. Parmar, Machine learning techniques for rainfall prediction: A review

[4] M. G. Hesham, F. Al-Turjman, and M. Zahid, A hybrid deep learning approach for optimizing hydrological prediction over sensor-cloud platforms

[5] Saranagata Kundu, Saroj Kr. Biswas, Deeksha Tripathi, Rahul Karmakar, Sounak Majumdar, Sudipta Mandal, A review on rainfall forecasting using ensemble learning techniques

[6] Dutta, P. S., Tahbilder, H. (2014), Prediction of rainfall using data mining technique over Assam