Rain Prediction in Sri Lanka

Contributors: Dhruv Kaushal, Arya Rao, Arjit Ghatta, Tanvi Pulipaka

1 Problem

In the heart of rural and suburban Sri Lanka, the majority of residents depend primarily on agriculture as a source of livelihood. Representing approximately 30% of the nation's workforce, agriculture is not merely a means of sustenance but a significant part of the Sri Lankan economy. The vitality of this industry depends on rainfall, making it a critical factor for crop yields.In the absence of adequate rainfall, crop growth and productivity are jeopardized. The consequences of this extend beyond economic concerns, impacting the lives of those reliant on the agricultural sector. To gain insight on potential challenges and rainfall patterns in Sri Lanka, our group has analyzed a collection of weather data for 30 cities in Sri Lanka from 2010 to 2023.

2 Related Works

In the paper "Rainfall Distributions in Sri Lanka in Time and Space", T.P. Burt and K.D.N. Weerasinghe conducted an analysis based on daily rainfall data in Sri Lanka This study talks about a relation among the rainfall features of Sri Lanka and farming, which is directly associated with our investigation on how climatic changes influences farming within a region. However, this research focuses on how rainfall patterns have changed through time and space, employing diverse analytical tools that help in understanding the complex concept of raining days. It had rainfall measures that included total rainfall, rain days, wet days and very wet days, noting limitations with t10 because rainfall is often skewed. They used the same statistical procedure to calculate weekly averages and determine the probabilities of obtaining 10 mm or more in a week as well as considering periods of dry and wet months using a markov chain model. These statistical techniques contribute to an informed interpretation of rainfall in Sri Lanka. The aim of this work is also to contribute to a more general understanding of the daily scale dynamics of precipitation processes and their associations with large climate factors, forming the basis for the prediction techniques that are applied in our project.

"Spatial variability of rainfall trend in Sri Lanka, 1989-2019: indication of climate change" was a study conducted by Niranga Alahachoon and Mahesh Edirsinghe that utilizes daily raining information from the CHIRPS database, and applies GIS strategies. The study utilizes Sen's slope estimator and the Mann–Kendall test to show an important rise in annual precipitation for every climate zone of Sri Lanka over that period. In this context, this understanding becomes relevant to our projects since it reveals the possible risks associated with flooding and droughts in some areas, thus providing essential information for the water resources management as

well as the disaster readiness. This points out the need to adopt proper dryness and wetness control procedures, which agrees with our efforts in boosting resilience under vulnerability to climate change.

3 Our Approach

The dataset used for this study contains 147,480 entries across 30 cities in Sri Lanka. There are numerous categories regarding weather data. However, not all of these columns are relevant and used in the model. We chose to use the following categories to predict `precipitation_sum`: `city`, `temperature_2m_max`, `temperature_2m_min`, `temperature_2m_mean`, `apparent_temperature_max`, `apparent_temperature_min`, `apparent_temperature_mean`, `shortwave_radiation_sum`, `windspeed_10m_max`, `windgusts_10m_max`, `winddirection_10m_dominant`, `latitude`, `longitude`, `elevation`, `precipitation_hours`, `weathercode`,`et0_fao_evapotranspiration`, `year`,and `month`. Furthermore, we removed the last ten days of data collected, as we used all factors from that time period excluding precipitation for our prediction.

The first step we took was to preprocess and clean the data. This included removing irrelevant columns in our dataset, checking for null or duplicate values, and extracting the date format to create a year and month column. To gain deeper insights into our dataset, we crafted data visualizations. For example, we created a graph displaying the average daily precipitation in Sri Lanka, utilizing data from 2010 to 2023. Furthermore, a heatmap was generated to depict the correlation among the selected factors. This provided us with a better understanding of what factors might be beneficial to investigate in their relationship with precipitation.

Following the data cleaning and visualization, we leveraged the refined dataset to construct various prediction models. Initially, we encoded the categorical variables, standardized the features, and split the data into training and testing sets. These categories included the weather code and the city. City was a categorical value that would not change and is essentially used as a bucket, and weather code is similar due to its nature being of a number which represented a group of weather conditions. Following this, we trained two distinct models on the training data: Linear Regression and Random Forest. The linear regression model aims to find the linear relationship between the selected features and the target variable `precipitation_sum`. Random forest comprises multiple decision trees that collectively make predictions using voting or averaging. The trained models were then employed to make predictions on the testing set.

4 Results

The two models that were used by our group to analyze the dataset were the Random Forest Regression and Linear Regression. Both of our models were trained based on the categories mentioned above (see Our Approach) in order to predict the target output of total precipitation sum to be predicted at each station. Due to a higher performance by the Random Forest Regression model we chose that for our primary analysis.

4.1 Random Forest Regression

Random Forest Regression: The Random Forest Regressor demonstrated better overall performance and predictive accuracy. By modeling non-linear relationships using decision trees, this approach delivered a low mean absolute error (MAE) of 0.788, mean squared error (MSE) of 9.231, root mean squared error (RMSE) of 3.038 and high R-squared (R2) value of 0.912. This indicates it explains approximately 91.2 percent of the observed variability in precipitation sums.

4.2 Linear Regression

The Linear Regression demonstrated a weaker overall performance and predictive accuracy. The model showed significantly higher error rates across all the metrics showing a mean absolute error (MAE) of 2.664, mean squared error (MSE) of 31.694, root mean squared error (RMSE) of 5.630, and high R-squared (R2) value of 0.697.

4.3 Predictions

After running our prediction based on splitting our test data and our training data we were able to output the precipitation sum per day. We grouped the overall averages by city and for the 20 percent of testing data which equates to 1.8 years to provide the average rainfall. (See Figure 1) We were able to accurately provide the average precipitation by day of the next almost two years. Following this calculation, we provided our model with data for the next 10 days from 2023. This data includes all of the categorical data that was used within our model, as well as, following the same steps for hot encoding the categorical variables of 'city' and 'weather code'. Utilizing this data, we used our model to predict the precipitation of the next 10 days. Our output displayed the average by city and day over the period of June 8th, 2023 to June 17th 2023.

5 Conclusion

Our study, which encompassed over a decade of extensive Sri Lankan weather data, has laid the groundwork for a predictive model with profound implications for Sri Lanka's agricultural sector. Through our model, users can gain insight into a predictive analysis for the rainfall within Sri Lanka based on the weather conditions. Potential implementations of our model lay within calculations of averages, trends, and predictions of weather conditions surrounding precipitation (features within our model). This predictive model has the potential for real-world utilization across various sectors in Sri Lanka. Most directly, the rainfall forecasts could provide farmers insight into making critical decisions about planting cycles, crop selection, water resource management, and harvesting timelines to boost yields. Additionally, disaster management units could incorporate these predictions into early warning systems for flooding, landslides, or drought. Overall our Random Forest model was able to provide in-depth predictions for Sri Lanka's rainfall which can lead to an improved decision-making process for a myriad of decisions.

6 Appendix

City	Average Rainfall
Athurugiriya	6.743186
Badulla	5.640986
Bentota	6.95223
Colombo	7.530518
Galle	5.472978
Gampaha	6.926449
Hambantota	3.265523
Hatton	7.615066
Jaffna	3.030826
Kalmunai	3.830339
Kalutara	6.91086
Kandy	6.292459
Kesbewa	7.441196
Kolonnawa	7.139093
Kurunegala	5.384287
Mabole	6.892312
Maharagama	6.675352
Mannar	3.435466
Matale	5.638719
Matara	5.894616
Moratuwa	6.966922
Mount Lavinia	6.90774
Negombo	6.878022
Oruwala	7.033505
Pothuhera	5.283162
Puttalam	5.207516
Ratnapura	7.792919
Sri Jayewardenepura Kotte	6.805623
Trincomalee	4.221901
Weligama	6.20572

Figure 1: Average Rainfall for 1.8 years (20 percent testing data)

7 Link to GitHub

https://github.com/avrao2/comp562

8 References

1. Alahacoon, N., & Edirisinghe, M. (2021, February 19). Spatial variability of rainfall trends in Sri Lanka from 1989 to 2019 as an indication of climate change. MDPI. https://www.mdpi.com/2220-9964/10/2/84

2. Burt, T. P., & Weerasinghe, K. D. N. (2014, September 26). Rainfall distributions in Sri Lanka in time and space: An analysis based on daily rainfall data. MDPI. https://www.mdpi.com/2225-1154/2/4/242

3. Rasul. (2023, July 12). Sri Lanka Weather Dataset. Kaggle. https://www.kaggle.com/datasets/rasulmah/sri-lanka-weather-dataset