AI-Based Sri Lankan Rainfall Prediction System

Jeyakumar.T (214096P)

1. Introduction

Rainfall prediction plays an important role in agriculture, disaster management, urban planning, and water resource management in Sri Lanka. Weather patterns vary significantly across districts and seasons, making short-term rainfall prediction a meaningful problem.

This project aims to build a machine learning model that predicts whether it will rain tomorrow for a given city using historical weather data.

The task is formulated as a binary classification problem, where:

- 1 → Rain Tomorrow
- 0 → No Rain Tomorrow

The project follows a structured machine learning workflow including preprocessing, leakage prevention, time-series splitting, model tuning, explainability, and deployment preparation.

2. Dataset Description

The dataset (srilanka_weather.csv) contains historical daily weather records including:

- time
- city
- latitude
- longitude
- precipitation_sum
- temperature features
- radiation features
- evapotranspiration
- sunrise / sunset
- other meteorological variables

The dataset spans multiple Sri Lankan cities across several years.

2.1 Data Source

The Sri Lanka Weather Dataset is a comprehensive collection of weather data for 30 prominent cities in Sri Lanka, covering the period from January 1, 2010, to January 1, 2023. The dataset offers a wide range of meteorological parameters, enabling detailed analysis and insights into the climate patterns of different regions in Sri Lanka.This dataset was sourced from Open-Meteo and simplemaps.

2.2 Target Variable Creation

The target variable rain_tomorrow was created by:

1. Creating a binary feature:
    - rain_today = 1 if precipitation ≥ 2.0mm
    - rain_today = 0 otherwise

2. Shifting per city:

    - Rain_tomorrow = rain_todayt+1

Note: Rows without a next-day value were removed.

3. Exploratory Data Analysis (EDA)

3.1 Dataset Overview

- Total rows: 147480
- Date range: 2010-01-01 – 2023-06-17
- Number of unique cities: 30

3.2 Missing Values

Missing values were inspected and quantified.

3.3 Rain Distribution (≥ 2mm Threshold)

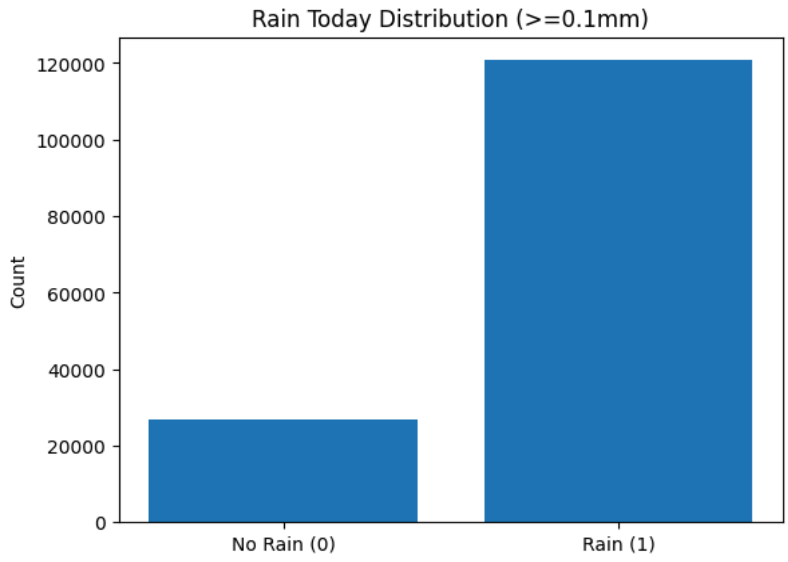

Figure 1: Rain Distribution bar chart

Target distribution:

- % No Rain Tomorrow: 55.75
- % Rain Tomorrow: 44.25

This indicates whether the dataset is balanced or slightly imbalanced.

3.4 Rainfall Over Time

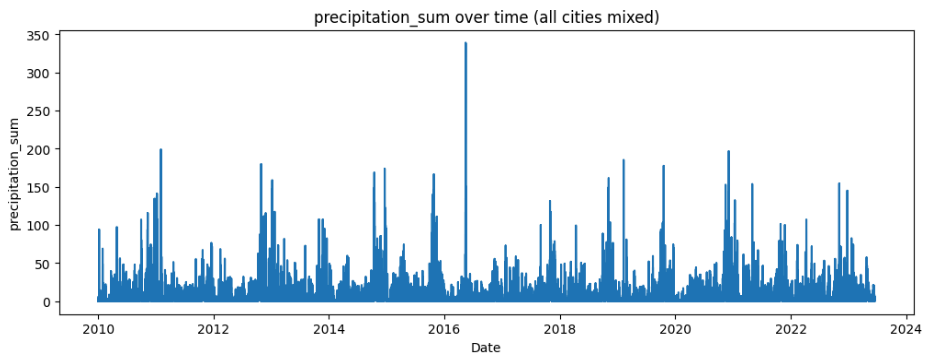

Figure 2 : Histogram of precipitation over time for all cities

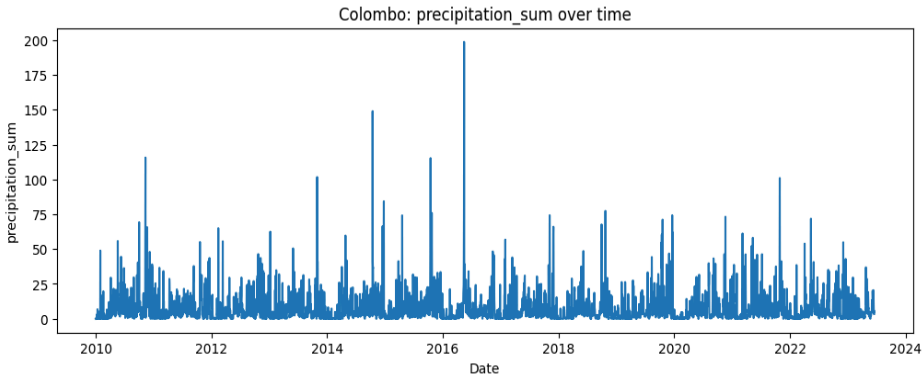

Figure 3 : Histogram of precipitation over time for colombo city

4. Leakage Prevention

To prevent data leakage, the following columns were removed before training:

- precipitation_sum
- rain_sum
- precipitation_hours
- rain_today
- weathercode
- snowfall_sum

These variables directly describe current rainfall and would artificially inflate performance.

5. Feature Engineering

Additional time-based features were extracted:

- month
- day_of_year
- year

City was one-hot encoded after splitting.

Non-numeric columns such as:

- time
- sunrise
- sunset

- country

were removed before training.

6. Data Splitting Strategy

A chronological split was performed using unique dates:

- First 80% of dates → Training
- Last 20% of dates → Testing

This ensures:

- No future data leakage
- Realistic forecasting scenario

Training and testing date ranges:

- Train: 2010-01-01 00:00:00 to 2020-10-06 00:00:00
- Test: 2020-10-07 00:00:00 to 2023-06-16 00:00:00

7. Baseline Model

A baseline XGBoost Classifier was trained using default parameters. Evaluation metrics like Accuracy, F1 Score and ROC-AUC were used.

Baseline Results

- Accuracy: 0.7126483553747033
- F1 Score: 0.726962237401727
- ROC-AUC: 0.7942014928784256

8. Hyperparameter Tuning

Hyperparameter tuning was performed using:

- RandomizedSearchCV
- TimeSeriesSplit (5 folds)
- F1 Score as optimization metric

Parameters tuned:

- n_estimators
- learning_rate
- max_depth
- subsample
- colsample_bytree
- min_child_weight
- gamma
- reg_alpha
- reg_lambda

9. Tuned Model Performance

Test Set Results

- Accuracy: 0.74774499830451
- F1 Score: 0.7797352915050484
- ROC-AUC: 0.8242516959933653

Comparison:

Metric	**Baseline**	**Tuned**
Accuracy	0.7126483553747033	0.74774499830451
F1	0.726962237401727	0.7797352915050484
ROC-AUC	0.7942014928784256	0.8242516959933653

Table 1 : Comparison Table for Matrices for baseline and tuned models

10. Threshold Optimization

Instead of using the default 0.5 threshold, multiple thresholds were evaluated from 0.1 to 0.9.

Best threshold for F1 score:

- Threshold: 0.30000000000000004

- F1: 0.8021712907117008
- Precision: 0.721768675381058
- Recall: 0.9027326919671841

This improves classification balance between precision and recall.

11. Explainable AI (SHAP Analysis)

SHAP was used to interpret model predictions.

11.1 Global Feature Importance

Top influential features may include:

- et0_fao_evapotranspiration
- shortwave_radiation_sum
- day_of_year
- temperature features
- humidity indicators

11.2 Single Prediction Explanation

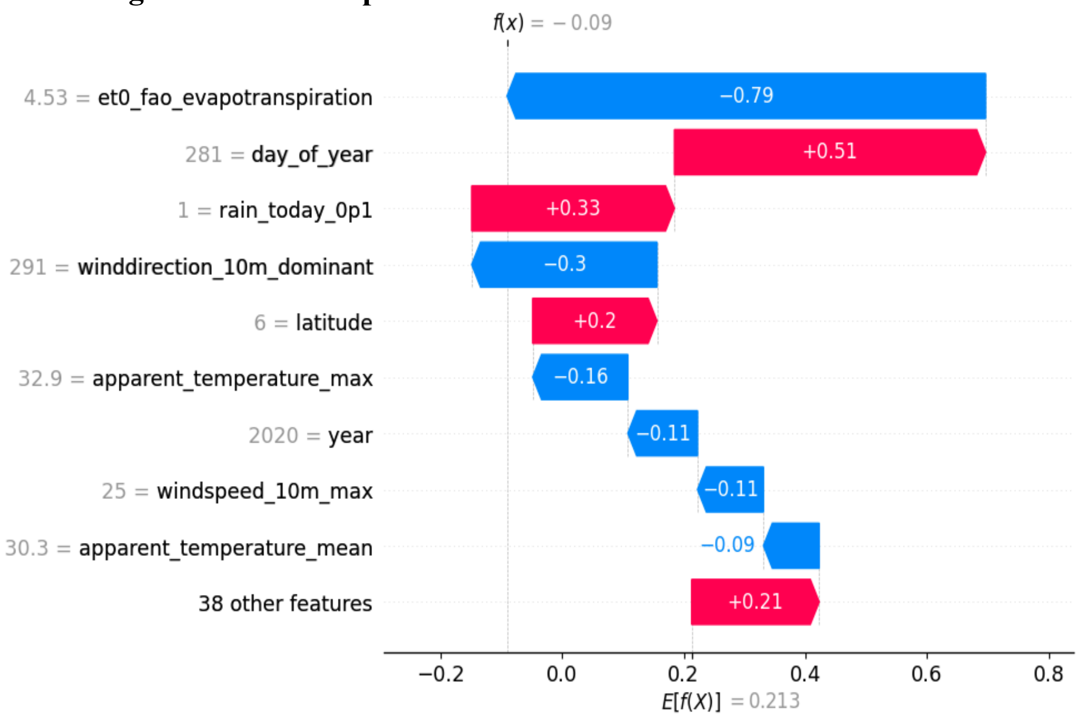

Figure 4 : Waterfall Plot for SHAP representation

This plot explains why the model predicted rain or no rain for a specific instance.

11.3 Dependence Plots

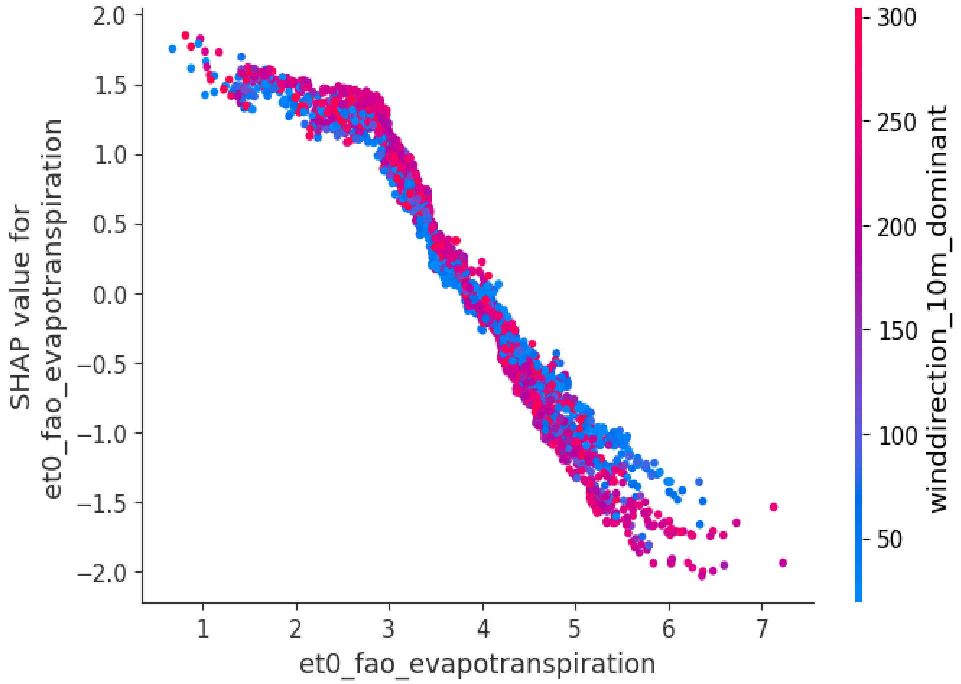

Figure 5 : SHAP Dependence Plots

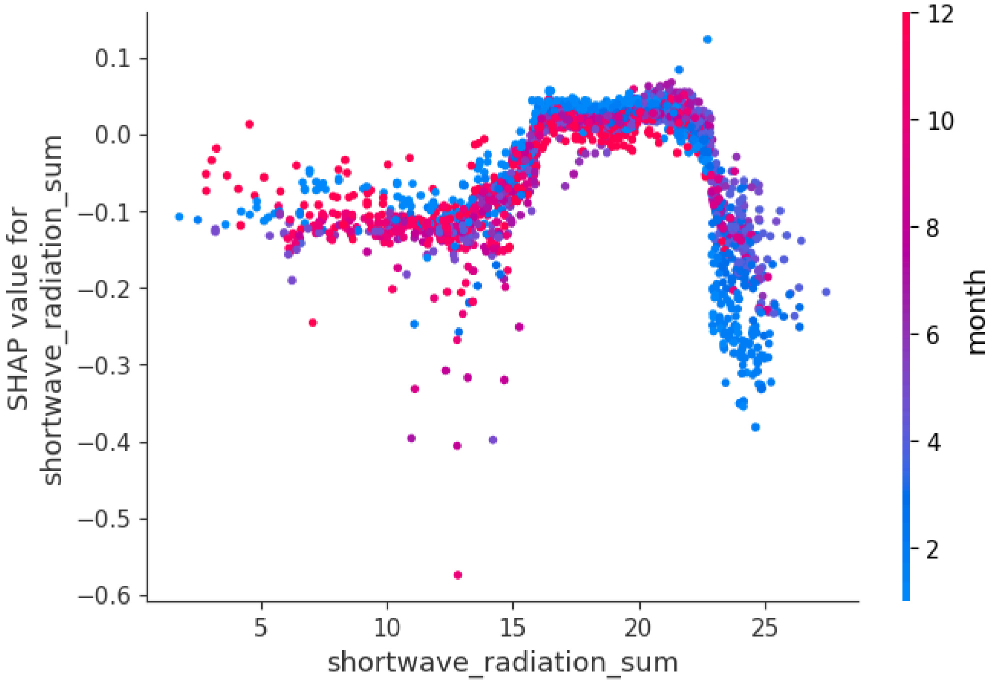

Figure 6 : SHAP Dependence Plots

These plots show how individual features influence rainfall probability.

12. Model Export

The final model was exported using:

- xgb_booster.json
- rain_artifact_meta.pkl

Saved components include:

- Trained booster
- Optimal threshold
- Feature column list
- Background dataset for SHAP

13. Critical Discussion

13.1 Limitations

- Dataset limited to available historical records
- No real-time atmospheric pressure maps
- No satellite imagery
- City-level granularity only

13.2 Data Challenges

- Potential missing observations
- Class imbalance in certain cities
- Seasonal variability

13.3 Real-World Use

The model predicts next-day rainfall probability and can support:

- Farmers
- Event planners
- Disaster management authorities

However, it should complement official meteorological forecasts.

14. Conclusion

This project demonstrates that XGBoost combined with time-aware splitting and SHAP explainability can effectively model next-day rainfall prediction in Sri Lanka.

The integration of threshold tuning and time-series validation ensures realistic and robust evaluation.