Genetic Programming and LSTM for Environmental Time Series Forecasting

Introduction

Over the past four weeks, I have been engaged in research and model development related to two major machine learning approaches—Genetic Programming (GP) and Long Short-Term Memory (LSTM) networks—with a focus on environmental and hydrological applications.

Literature Review

My work began with a thorough literature review on Genetic Programming and its applications in Water Resources Engineering and Hydrology. I studied the role of GP in modeling complex nonlinear processes like rainfall-runoff relationships, groundwater level prediction, and environmental risk mapping. The focus was on how GP evolves interpretable mathematical expressions that can model physical phenomena, especially where empirical relationships are hard to derive. I also reviewed how GP integrates with environmental monitoring and remote sensing systems, highlighting its use in symbolic regression and spatial data modeling.

Work Done So Far

Subsequently, I was assigned to study LSTM-based time series forecasting using two univariate datasets from IMD: temperature and precipitation. This involved preprocessing the data, such as timestamp formatting, normalization, and reshaping it into supervised learning format. Although the datasets had no missing values, I explored gap-filling techniques using interpolation as a precaution to maintain temporal consistency.

I have started building forecasting models in two phases: single-step (predicting the immediate next value) and multi-step (predicting several future values). I am experimenting with different LSTM architectures including Vanilla, Stacked, and Bidirectional LSTM models. Separate models are being trained for temperature and precipitation, since temperature follows more regular seasonal trends, while precipitation is more erratic and spike-driven.

Proposed Work

Model development is still in progress. The next steps include model evaluation using metrics like MAE and RMSE, inverse-scaling the predictions to original units, and comparing performance across LSTM variants.

Conclusions

This 4-week phase has provided strong foundations in both symbolic modeling with GP and sequence modeling with LSTM. The work is progressing steadily toward a deployable forecasting system for environmental applications.