Regression Analysis of Rainy Weather Dataset: Using Different Machine & Deep Learning Models

Misbahul Sheikh
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
misbahul.cse.20200204039@aust.edu

Naushin Mamun
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
naushin.cse.20200204010@aust.edu

Nafisa Tasnim Neha
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
nafisa.cse.20200204020@aust.edu

***Abstract*—This paper conducts a regression analysis on rainy weather data using various machine learning and deep learning models to predict continuous variables related to rainfall. Five models were tested: Support Vector Regression (SVR), Linear Regression (LR), Random Forest Regression (RFR), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNN). The models were thoroughly trained and validated for robustness. RFR showed superior accuracy. Deep learning models, LSTM and CNN, achieved the lowest mean absolute errors of (0.25) and (0.24), and mean squared errors of (0.12) and (0.13), capturing temporal and spatial dependencies effectively. This study demonstrates the superior performance of deep learning models in predicting meteorological data.**

***Keywords*—Regression analysis, rainy weather dataset, SVR, LR, RFR, LSTM, CNN, machine learning, deep learning.**

I. Motivation

Accurate prediction of rainfall is crucial for various sectors, including agriculture, water resource management, disaster preparedness, and urban planning. Traditional weather forecasting methods often fall short in capturing the intricate patterns and dependencies in meteorological data. With the advent of advanced machine learning and deep learning techniques, there exists a potential to significantly enhance the accuracy and reliability of rainfall predictions.

This paper aims to explore and compare different regression models, including both traditional machine learning algorithms and deep learning approaches. By leveraging the strengths of models such as Support Vector Regression (SVR), Linear Regression (LR), Random Forest Regression (RFR), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNN), this study seeks to provide a robust framework for improved weather forecasting.

The motivation is to bridge the gap between theoretical advancements in machine learning and their practical application in meteorology, ultimately contributing to better-informed decision-making and mitigation strategies in response to weather-related challenges.

II. Literature Review

Some past works related to this topic were studied for this analysis.

A paper [1] analyzes the proposed model for rainfall prediction using linear regression. The model was trained with historical rainfall data and other weather-related information from Chennai. The results showed that the linear regression model effectively gives us the results. The linear regression model of that paper gives mean absolute error and R-square value of (0.123) and (0.946) respectively.

Another [2] study focuses on predicting rainfall using multiple linear regression (MLR) models. The dataset comprises 30 years of regional rainfall data from Rajshahi, Bangladesh, along with other climatic factors such as precipitation, cloud cover, average temperature, and vapor pressure. They finds a very small error in their regression model with mean absolute error, mean square error, root mean square error and R-square value of (0.08), (0.23), (0.20) and (0.823) respectively.

The paper [3] investigates rainfall prediction using various machine learning models. The dataset, includes rainfall data for 36 subdivisions in India. The models were evaluated using the R² score. Among the models, MLR achieved the highest R² score (0.9950), followed by Lasso Regression (0.9910), XGBoost (0.9819), and RFR (0.9737). The study highlights the superiority of MLR for this dataset, emphasizing its effectiveness in capturing rainfall patterns despite the inherent non-linearity in climatic data .

Another [4] paper describes a statistical forecasting model for summer monsoon rainfall in Bangladesh. The literature review notes the shift to statistical models for better accuracy. Significant relationships between these predictors and monsoon rainfall were found from 1961 to 2007. The model's performance showed an RMSE of (8.13), a Heidke skill score of (0.37), and a correlation of (0.74) between predicted and observed rainfall. The BIAS, as a percentage of the long period average (LPA), was -0.85%. The model had a hit score of 58%.

III. Methodology

A. Data Collection

The Data have been collected from kaggle.com website. This dataset contains total 3271 atmospheric data with 2 classes of rain today: Yes (25.96%) & No (74.04%). There are total 20 attribute columns including the target column in the dataset.

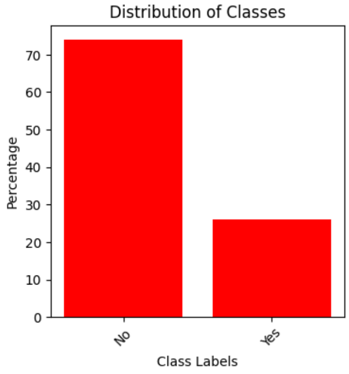

Fig. 1. Class Distribution Graph

B. Data Preprocessing

We performed several tasks in the pre-processing stage. Text data was converted into numeric values to ensure compatibility with machine learning algorithms, and null values were removed to maintain data integrity and prevent inaccuracies in predictions. These steps are crucial for preparing the data for effective regression analysis.

C. Regression Analysis

Regression analysis is a fundamental statistical technique used to model the relationship between a dependent variable (target variable) and one or more independent variables (predictors).

D. Data Splitting

We have split the data into train and test sets and the ratio was (train: test: validation) = (8:2:2). We kept the validation set the same as the testing set and used 100 epochs.

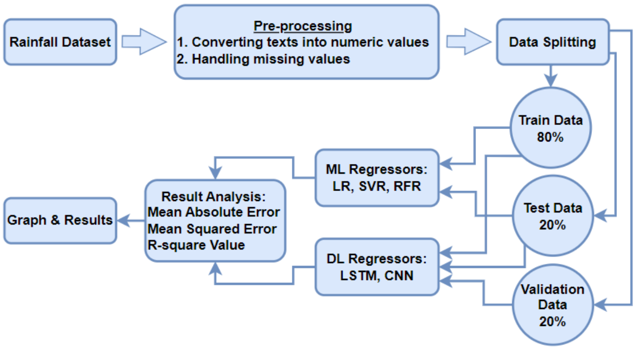
Rainfall Dataset
Pre-processing
1. Converting texts into numeric values
2. Handling missing values
Data Splitting
Train Data 80%
Test Data 20%
Validation Data 20%
ML Regressors: LR, SVR, RFR
DL Regressors: LSTM, CNN
Result Analysis: Mean Absolute Error Mean Squared Error R-square Value
Graph & Results

Fig. 2. Analysis Model

IV. Result Analysis

The performance of our rainy weather dataset was evaluated using key metrics such as mae, mse, and r2 value.

Regressor	MAE	MSE	R2
SVR	0.28	0.18	0.05
LR	0.27	0.13	0.32
RFR	0.25	0.13	0.30
LSTM	0.25	0.12	0.35
CNN	0.24	0.13	0.34

TABLE I
PERFORMANCE METRICS

Deep learning models – LSTM and CNN provides us a better result with minimum errors and maximum r-square values among all the models. SVR analysis gives us the poorest results with maximum errors.

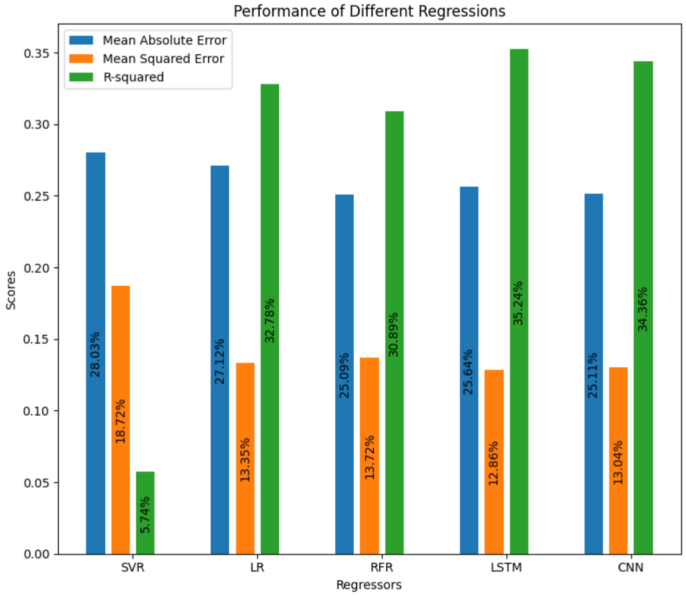

Fig. 3. Performance Graph

References

[1] J.Refonaa, M.Lakshmi, Raza Abbas, Mohammad Raziullha, "Rainfall Prediction using Regression Model," 2019 International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-2S3.

[2] MAI Navid, NH Niloy, "Multiple Linear Regressions for Predicting Rainfall for Bangladesh," 2018 Communications. Vol. 6, No. 1, 2018, pp. 1-4. doi: 10.11648/j.com.20180601.11.

[3] Santosh Singh, Vishal Pandey, Abhishek Singh, "Rainfall Prediction Using Machine Learning," Thakur College of Science & Commerce, Thakur Village, Kandivali East, Mumbai 400101, Maharashtra, India. 2024 IJCRT, Volume 12, Issue 3 March 2024, ISSN: 2320-2882.

[4] MD MIZANUR RAHMAN, RAFIUDDIN, MD MAHBUB ALAM, "Seasonal forecasting of Bangladesh summer monsoon rainfall using simple multiple regression model," 2013 SpringerLink, Volume 122, pages 551–558, (2013).