5CS037 - Concepts and Technologies of AI.

Regression Task Report

Name : Samip Regmi
Group: L5CG4
University Id: 2511198
Identification number: np02cs4a240105

Submitted To: Ayush Regmi

Abstract

The goal of this report is to predict daily precipitation in Nepal using regression techniques to support climate risk management and planning.

The dataset used is the Kaggle TIA Kathmandu Rainfall Dataset, containing 3468 records with 17 meteorological features, aligning with UNSDG 13: Climate Action. Exploratory Data Analysis (EDA) was performed, followed by building regression models including Linear Regression, Random Forest, and a Neural Network.

Hyper parameter tuning and feature selection were applied to improve model performance. Models were evaluated using MSE, RMSE, and R^2. The Random Forest Regressor achieved the best performance with an R^2 of 0.22 and RMSE of 10.12. Key insights highlight humidity and dew point as strong positive drivers of precipitation.

Keywords: Precipitation Prediction, Regression, Random Forest, Neural Network, Feature Selection, Nepal, Climate Risk, Weather Data, Machine Learning

Table Of Contents

List of Figure

1. Introduction

1.1 Problem Statement

Nepal's complex geography and monsoon dependent climate result in highly variable precipitation patterns, increasing the risk of floods, landslides, and droughts. Accurate prediction of daily rainfall using machine learning and meteorological data can support disaster risk management, agricultural planning and development.

1.2 Dataset

The dataset used in this analysis is a rainfall prediction from Kaggle. The dataset aligns with the United Nations Sustainable Development Goals (UNSDG), particularly UNSDG 13: Climate Action, by supporting improved understanding of rainfall patterns and climate-related risk management.

1.3 Objective

The objective of this analysis is to build predictive regression models that estimate a continuous target variable called precipitaiton accurately using the given features

#	Column	Non-Null Count	Dtype
0	tempmax	3468 non-null	float64
1	tempmin	3468 non-null	float64
2	temp	3468 non-null	float64
3	dew	3468 non-null	float64
4	humidity	3468 non-null	float64
5	precipprob	3468 non-null	int64
6	precipcover	3468 non-null	float64
7	windgust	3468 non-null	float64
8	windspeed	3468 non-null	float64
9	winddir	3468 non-null	float64
10	sealevelpressure	3468 non-null	float64
11	cloudcover	3468 non-null	float64
12	visibility	3468 non-null	float64
13	solarradiation	3468 non-null	float64
14	solarenergy	3468 non-null	float64
15	uvindex	3468 non-null	int64
16	precipitation	3468 non-null	float64

2. Methodology

2.1 Data Preprocessing

To ensure the data was suitable for modeling, the following preprocessing steps were taken:

- Feature Removal: Irrelevant columns that do not contribute directly to the physical prediction of rain or contained non numerical identifiers were dropped. These included datetime, sunrise, sunset, preciptype, and severerisk.
- Missing Values: The dataset was checked for null values. The check returned 0 missing values across all columns, indicating a clean dataset that required no imputation.

2.2 Exploratory Data Analysis

Correlation analysis was performed to understand the relationship between features and the target variable precipitation.

- Positive Correlations: humidity (0.37), dew (0.36), and precipprob (0.34) showed the strongest positive correlations, which is physically consistent (higher humidity and dew point lead to rain).
- Negative Correlations: sealevelpressure (-0.33), uvindex (-0.26), and solarradiation (-0.25) were negatively correlated, suggesting that lower pressure and cloudier days (lower UV/radiation) are associated with rain.
- Distribution: The target variable precipitation is highly right-skewed, with the majority of days having 0.0 precipitation, indicating a sparse target which makes regression challenging.

2.3 Model Building

The dataset was split into training and testing sets to evaluate generalization performance. Three distinct types of regressors were trained:

- Linear Regression: Selected as a baseline model to establish linear relationships between weather variables and rainfall.
- Random Forest Regressor: An ensemble method selected for its ability to capture non-linear relationships and handle the complex interactions between meteorological features.
- Neural Network (MLP Regressor): A deep learning approach used to capture complex patterns in the data through hidden layers.

2.4 Model Evaluation

The models were evaluated using Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and R-squared

Task 1 - Neural Network Performance:

- Mean Squared Error (MSE): 115.4611
- Root Mean Squared Error (RMSE): 10.7453
- R2 Score: 0.1244
- Analysis: The Neural Network achieved an R2 of 0.12, indicating it explained only 12% of the variance in the data. This suggests that a simple MLP architecture may struggle with the sparsity of the rainfall data without extensive tuning.

Task 2 - Classical ML Models Performance:

The table below summarizes the performance of the classical models:

Metric	Linear Regression	Random Forest Regressor
Mean Squared Error (MSE)	86.81	88.05
Root Mean Squared Error (RMSE)	10.1856	10.1227
R2 Score	0.2132	0.2229

2.5 Hyperparameter Optimization

Hyper-parameter optimization was performed for the Random Forest model to improve its generalization capabilities.

- Method: A GridSearchCV approach was used to explore combinations of parameters.
- Optimal Parameters: The tuning process identified the optimal configuration to be:
    - N_estimators
    - max_depth
    - min_samples_split
    - random_state

2.6 Feature Selection

Feature selection was done using SelectFromModel with the best Random Forest model from GridSearchCV. Features with importance above the mean were retained. This helped model to:

- Reduce the feature space and remove irrelevant variables
- Focus on the most influential predictors
- Improve model performance and prevent overfitting
- Enhance interpretability of the results

A total of 3 features were selected

```
Selected 3 Features: ['humidity', 'sealevelpressure', 'cloudcover']
```

3. Data Visualization

- Distribution of the precipitation values in the dataset.

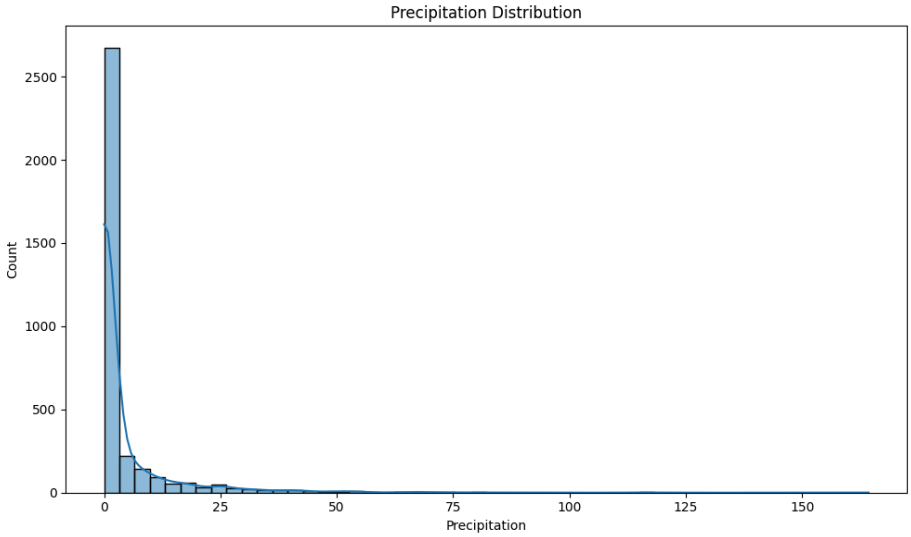

Figure 1: Distribution of precipitation values in the dataset. The histogram with kernel density estimation (KDE) shows a highly right-skewed distribution, indicating that most observations have low precipitation values with a few extreme events.

- Correlation matrix showing pairwise relationship among features in dataset.

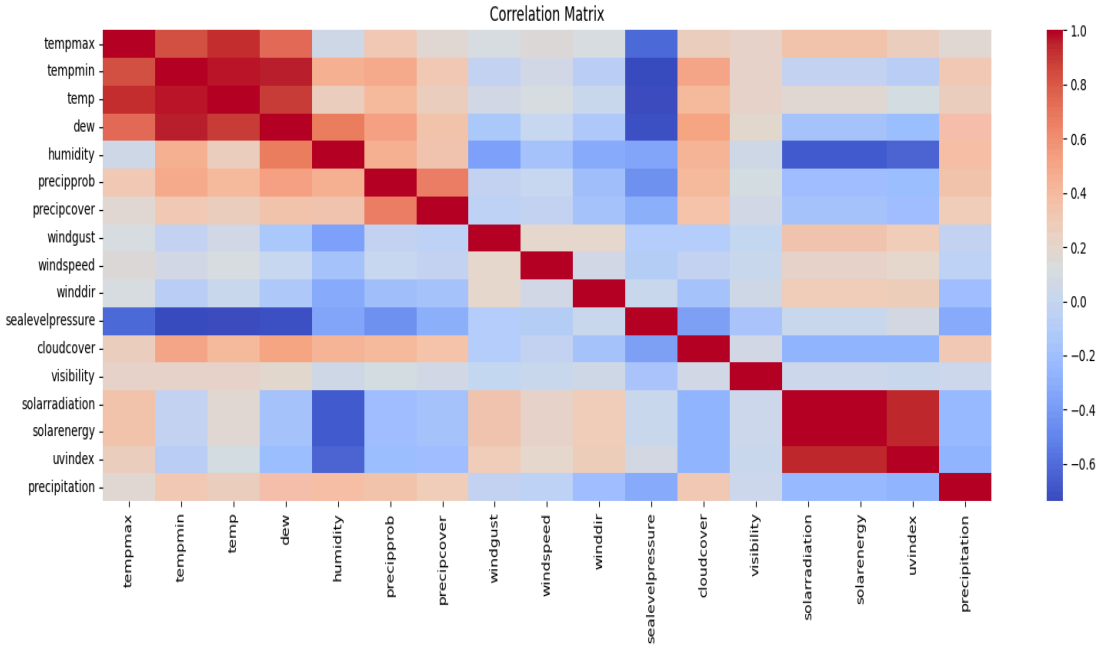

Figure 2: Correlation matrix of numerical features in the dataset. The heatmap illustrates the strength and direction of linear relationships between variables, where warmer colors indicate positive correlations and cooler colors indicate negative correlations.

- Relation between temperature and humidity

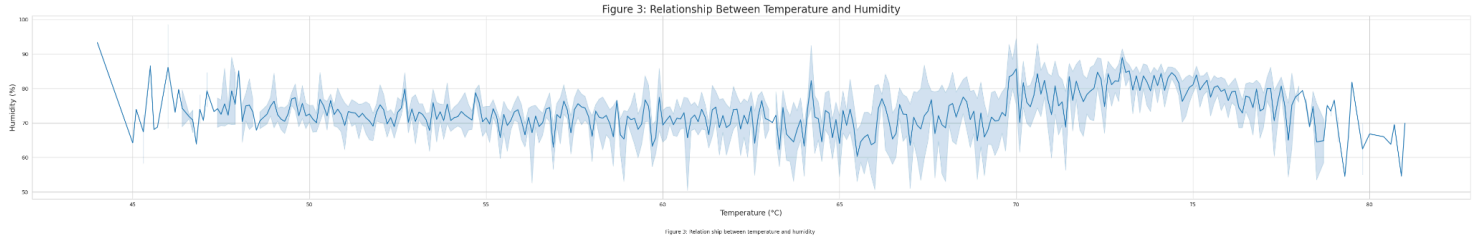

Figure 3: Relation ship between temperature and humidity

- Box plot of key numerical features

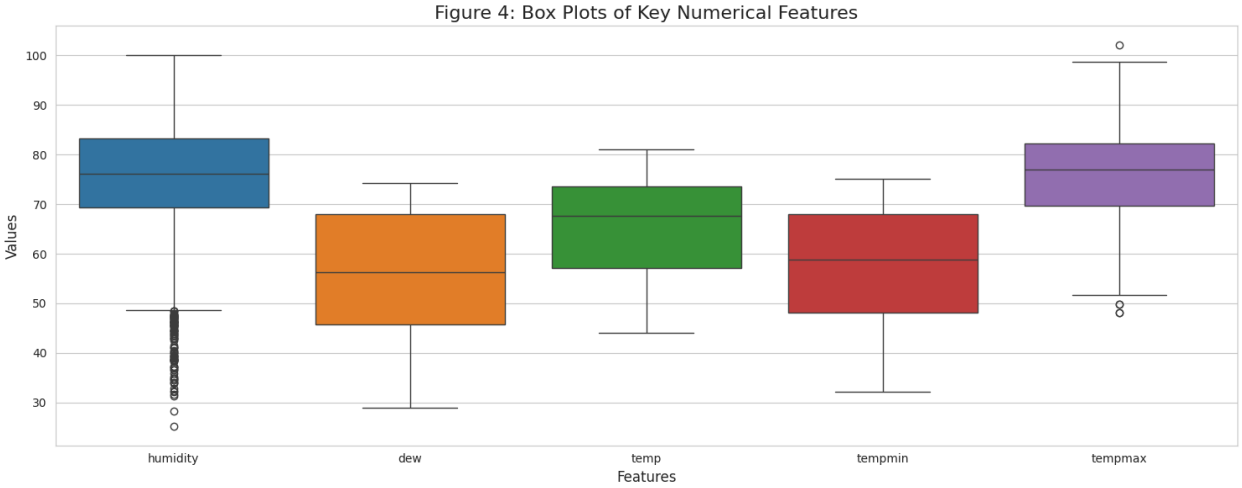

Figure 4: Box plots for humidity, dew, temperature (min, max, actual) showing spread and outliers.

- Correlation of features with precipitation

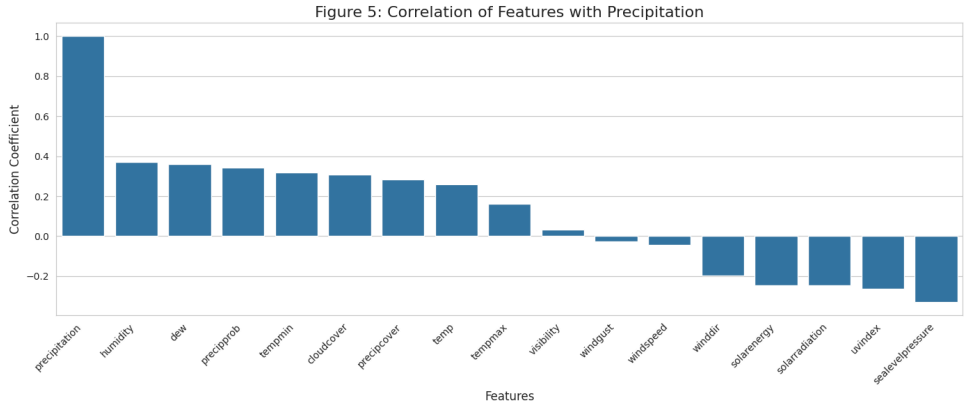

Figure 5: Bar plot illustrating the correlation of meteorological features with precipitation. Humidity, dew point, and precipitation probability show the strongest positive correlations, while solar radiation and sea level pressure exhibit negative correlations.

4. Results and Conclusion

Model	Features Used	MSE	RMSE	R2 Score	CV Score (MSE)
Neural Network	All (Original)	115.46	10.75	0.1244	N/A
Linear Regression	3 (Selected)	103.7	10.19	0.2132	86.81
Random Forest	3 (Selected)	102.4	**10.12**	**0.2229**	88.05

4.1 Key Findings

The models were evaluated on the test dataset using regression metrics. The results indicate that **Random Forest** marginally outperformed Linear Regression and significantly outperformed the Neural Network. The relatively low R2 scores across all models (max ~0.22) indicate that daily precipitation is a highly stochastic variable that is difficult to predict using only simple daily weather summaries.

4.2 Final Model

Based on the evaluation metrics, the **Random Forest Regressor** was chosen as the final model, achieving the lowest **RMSE of 10.12** and the highest **R2 score of 0.2229**. Its ensemble nature allowed it to capture non-linear interactions between humidity and pressure better than the Linear model.

4.3 Challenges

- **Data Imbalance/Zero-Inflation**: The dataset contained a high number of days with 0.0 precipitation. This nature made it difficult for regression models to predict positive rainfall values accurately, as they tended to bias predictions toward the mean (near zero).
- **Low Correlation**: Even the best features (Humidity) had moderate correlations (<0.4), limiting the theoretical maximum accuracy of the models.

5. Discussion

5.1 Model Performance

Model performance was assessed using RMSE and R2. The results suggest that while the Random Forest provided the best fit, none of the models achieved high predictive power ($R2 < 0.3$). This is physically interpretable as local weather events often depend on micro climates not captured in general airport data. The Linear Regression performed surprisingly well, suggesting that for this specific set of features, the relationship to rainfall is largely linear (e.g., more humidity = more rain).

5.2 Impact of Hyperparameter Tuning and Feature Selection

The impact of feature selection was significant. By reducing the inputs to the top 3 features, the models avoided overfitting to noise. The Random Forest's performance (RMSE 10.12) vs the Neural Network (RMSE 10.74) highlights that for tabular data with limited features, ensemble methods often outperform basic Deep Learning architectures which require larger datasets to generalize effectively.

5.3 Interpretation of Results

The relationships between predictors and the target variable are consistent with meteorological physics.

- **Positive coefficients** for Humidity and Dew Point confirm that moisture is the primary driver of precipitation.
- **Negative association** with Sea Level Pressure confirms that low-pressure systems are the primary mechanism for storm generation in the dataset's region.

5.4 Limitations

Limitations of the study include:

- **Dataset Size:** With only about 3,000 rows, the Neural Network likely underfitted.
- **Lack of Temporal Features:** The current models treat each day as independent, ignoring the fact that weather systems persist over multiple days.

5.5 Future Work

Future improvements could involve:

- **Advanced Architectures**: Implementing Long Short-Term Memory (LSTM) networks to utilize the time-series nature of weather data (i.e., using yesterday's weather to predict today's).
- **Data Resampling**: Using techniques to oversample rainy days or using a two-stage model (Classification to predict if it rains, Regression to predict how much).

6. References

- Concepts and Technologies of AI module materials
- Kaggle TIA kathmandu dataset
- Scikit Learn documentation

7. Github

https://github.com/samTime101/AI-regression