Herald College, Kathmandu

Concepts and Technologies of AI

5CS037

SDG Goal 13 — Climate Action: Climate Data Regression Analysis

Module: 5CS037 — Concepts and Technologies of AI

Assessment: Final Portfolio Project

Student Name: Ishuv Giri

Student ID: 2461987

Institution: Herald College Kathmandu (University of Wolverhampton)

Date: February 2026

Contents

1. Introduction

1.1 Background

One of the most urgent international concerns in the 21 st century is climate change, right? The United Nations Sustainable Development Goal 13 Climate Action is personal in terms of cause because it requires the urgent action to combat climate change and its effects. The historical climate trends are also encouraged to be determined by digging into the past with the use of data-driven approaches, which will help to predict the future trends and policies to be made. In this project, machine learning is introduced into the fray to examine historical records of climate conditions by Global Historical Climatology Network (GHCN) with the view to forecasting the value of climate measurements in terms of both time and observational characteristics. In the manner I am demonstrating by constructing and benchmarking a series of regression models, I am demonstrating how AI can, in fact, be used to address climate-related issues as in SDG 13

1.2 Aim and Objectives

Goal: Develop and experiment machine-learning univariate regression models that are capable of predicting climate observation values using historic weather data in weather stations.

Objectives:

- Complete exploratory data analysis of GHCN daily climatic records since 1901.

- Training and building a Multi-layer Perceptron (MLP) neural network regressor.

- Train two standard machine-learned models, the Random Forest Regressor and the Linear Regression to compare them directly.

- Hyperparameters of the fine-tuning model.

- Use feature-selection technique to winnow variables that are the most predictive. - Compare now all the models to determine the best performing.

1.3 SDG 13 Relevance

The analysis of climatic data supports SDG 13 in the following ways:

- Provides an empirical understanding of how climate has old fashioned.

- Helps to develop predictive models of most important climate variables such as temperature and precipitation.

- Provides instruments of analysis that can guide the thinking on adaptations and mitigation to climate change.

- D demonstrates how AI and machine learning can increase climate science and environmental measures.

2. Dataset Description and Preprocessing

2.1 Dataset Overview

In this project, the dataset used is ghcndaily1901kaggle.csv which is available in GHCN Daily collection. It includes the daily weather observations of weather stations located throughout the globe all done in 1901.

Property	Details
File	ghcn_daily_1901_kaggle.csv
Original Rows	50,000
Columns	9 (station_id, date, element, value, mflag, qflag, sflag, obstime, unit)
Date Format	ISO format (1901-01-01)
Climate Elements	TMAX, TMIN, TAVG, PRCP, SNWD, SNOW, TOBS, and

	others
Target Variable	value — the recorded climate measurement

2.2 Data Cleaning

The following preprocessing steps were applied:

- Duplicate Removal Slice 0 duplicate rows - already was cleaned here. - **Missing Value Analysis: Massive missingness: in both mflag (91.7), qflag (98.8), and obstime (96.9). Those were columns that were not necessary to be dropped.

- Quality Flag Filtering: Rows that are marked by *qflag* as unreliable are dropped, and 49,381 rows (619 dropped).

- Element Filtering: Removed the four main climate elements and left only the four major elements of TMAX, TMIN, TAVG, and PRCP which reduced the rows to 37,026.

- Outlier Removal: Trim above 1 st and 99 st percentile, resulting in 36,306 observations.

- Column Dropping: There is a drop of non essential columns (mflag, sflag, qflag, obstime).

Final cleaned dataset: 36,306 rows × 5 columns.

2.3 Feature Engineering

Carried out the extraction of temporal features of the date column to increase the feature space:

Feature	Description
year	Year of observation (1901)
month	Month of observation (1–12)
day	Day of month
day_of_year	Day of year (1–365)
season	Season derived from month (Winter, Spring, Summer, Autumn)

The variables that are categorical (stationid, element, unit, season) were encoded using the LabelEncoder to match them. The last set of features consisted of 8 engineered features.

3. Exploratory Data Analysis

3.1 Distribution Analysis

I verified the values of climate values using box plots and histograms. There is a great variety of data, indicating the differences between climate measures (temperature in tenths of degrees Celsius, precipitation in tenths of millimetres). It is fairly symmetric about zero, and negative and positive values are far apart spread widespread largely due to temperature records.

3.2 Correlation Analysis

A correlation heatmap pointed at the relations between numeric features and the target (value). Key takeaways:

- The highest correlation was between Element type and climate value; the various types of elements (TMAX, TMIN, PRCP) have varied ranges of measurements as part of their nature.

- Stationid was moderately correlated and this indicates geographic variation.

- There were positive but weak temporal features, such as day, dayofyear, which inherently had a seasonal relationship with temperature.

- The linear correlation of the year, month and season with the target was insignificant.

3.3 Temporal Patterns

- Montly tendencies: There was a variability in the average climate values on a monthly basis that exhibited normal seasonal variations in temperature patterns.

- Seasonal variation: Seasonal box plots showed that the trends expected were as follows, larger values on warmer days, and smaller on colder ones.

- Element distribution: TMAX, TMIN, and PRCP had different values ranges, which demonstrated the relevance of the element type as an indicator.

3.4 Key EDA Insights

It is the most informative variable and element feature, as in defining the scale and range of measurements it is a fundamental difference. Geographic location (stationid) has much variation based on the variations in climate zone. The availability of seasonal patterns is there but with a fairly low t-value, indicating the utility of non-linear models. The quality of data is identified as

typically good post filtering and no systematic biases were found.Data quality is generally good after filtering, with no systematic biases detected.

4. Model Development

There are three regression models that were generated and tested, which included: an MLP neural network, a Random Forest ensemble, and a Linear Regression baseline.

4.1 MLP Regressor (Neural Network)

The first model is MLP regressor, which is a type of a neural network (Neural Network). Justification Architecture The Multi-layer Perceptron (MLP) was selected because it has the capability to learn non-linear relationships, which are characteristic of climate data. Climate trends are not very straightforward as many variables interact with each other and thus cannot be understood by a simple linear model. Value 1 10.4 -86 This fixed parameter represents the typical average value of a variable.<|human|>Parameter 1 Value Justification This fixed parameter is the average value of a variable. Hidden Layers (128, 64, 32) 3-layer progressive design of hierarchical feature learning. Efficiently computed, activation-free, does not vanish Escaping activation gradient Solver Adaptive rates of convergence. Learning Rate= 0.001=Standard starting rate Adam optimiser. Max Iterations 500 Sufficient convergence with early stopping. Early Stopping True In early stopping, the validation loss is used to avoid overfitting. Batch Size 64 Trades off gradient quality and training speed.

Parameter	Value	Justification
Hidden Layers	(128, 64, 32)	Three-layer progressive architecture for hierarchical feature learning
Activation	ReLU	Efficient computation, avoids vanishing gradient
Solver	Adam	Adaptive learning rates for stable convergence
Learning Rate	0.001	Standard starting rate for Adam optimiser
Max Iterations	500	Sufficient for convergence with early stopping
Early Stopping	True	Prevents overfitting by monitoring validation loss
Batch Size	64	Balances gradient quality with training speed

Results:

Metric	Value
MSE	2826.85
RMSE	53.17
MAE	31.92
R^2	0.4087
Training Iterations	101

4.2 Random Forest Regressor

Model Justification:

Against common sense, Random Forest was selected as it is a robust, classical baseline, that is inherently non-linear and non-interactive in its nature. It is also bagging-robust, contains over-fitting in check, and imparts feature importance- a very pleasant explanation hook to us students.Initial Configuration: 100 trees, max_depth=15, min_samples_split=5, min_samples_leaf=2.

Results:

Metric	Value
MSE	1385.47
RMSE	37.22
MAE	21.51
R^2	0.7102

The R.F. is much more successful than the MLP and it has the R 2 = 0.7102 thus, it predicts 71% of the variance. The importance of features indicates that stationid (0.58) and element (0.26) are the most important features, then there are dayofyear (0.08) and day (0.08). Temporal characteristics such as year, month and season do little.

4.3 Linear Regression

Model Justification:

Linear Regression forms a baseline upon which we check its poor performance and the fact that it is not linear-friendly. It assists us in noting the necessity of more sophisticated models and yet enabling us to interpret the coefficients.

Results:

Metric	Value
MSE	4329.89
RMSE	65.80
MAE	40.44
R^2	0.0943
Intercept	5.2716

The data is obviously non-linear as Linear Regression only explains 9 percent of the variance (R 2 =0.0943). Coefficient Analysis: the largest negative coefficient is of element (0 48.18) and is

positive (35.06), stationid is moderate (4.28), day and dayofyear are tiny positives (0.92 each), and year, season, month are close to zero.

Coefficient Analysis:

- element had the largest negative coefficient (-48.18), indicating its strong influence on predictions.
- unit had a positive coefficient (35.06), reflecting the measurement scale differences.
- station_id contributed moderately (4.28).
- day and day_of_year had small positive coefficients (0.92 each).
- year, season, and month had zero or near-zero coefficients.

Actual vs Predicted Plots:

The scatter plots comparing actual vs predicted values across all three models visually confirmed the performance hierarchy. The Random Forest predictions clustered closest to the ideal diagonal line, while Linear Regression predictions showed horizontal banding — a clear sign of underfitting due to the model's inability to capture non-linear patterns.

5. Hyperparameter Optimization

5.1 Random Forest — GridSearchCV

Hyperparameter Grid:

Parameter	Values Tested
n_estimators	[50, 100, 200]
max_depth	[10, 15, 20]
min_samples_split	[2, 5, 10]

Method: 5-fold cross-validation with negative MSE scoring. Total: 27 combinations × 5 folds = 135 fits.

Best Parameters:

- n_estimators: 200
- max_depth: 20
- min_samples_split: 5

Best CV RMSE: 38.72

A deeper, many-tree setups capture the patterns in the climate, and with 20 reaches as deep as the tuned Random Forest (RMSE of CV 38.72), it can be seen that the deeper the setups, the more precise the patterns are represented by them

5.2 Linear Regression — Cross-Validation

No hyperparameters to optimise, 5 -fold CV performed generalisation. CV RMSE: 64.71. Like with held-out data the CV RMSE remained relatively steady at 64-66, which highlights the fact that a linear model is not fitting

6. Feature Selection

6.1 Method

Two filters (SelectKBest) were used to choose the best 4 out of 8 features (50 34). f regression (F-statistic): verifies linear dependency. mutual-info-regression: encompasses the linear and non-linear connections.

6.2 Results

f_regression Scores:

Feature	F-Score	Selected
element	1182.80	Yes
unit	426.42	Yes
day_of_year	19.77	Yes
day	19.77	Yes

mutual_info_regression Scores:

Feature	MI-Score	Selected
element	0.6146	Yes
unit	0.5598	Yes
station_id	0.3591	Yes
day_of_year	0.0265	Yes

6.3 Analysis

There is an agreement in both approaches that the largest signals are the element and unit. Markedly, however, mutual information, which substitutes day of the F-test, finds stationid to be important (MI = 0.36). The difference indicates non-linear dependence among location and climate that F-statistics used linearly do not capture. Final selected features (MI): stationid, element, unit, dayofyear - four of them reduce the feature set half and retain the most informative variables.

7. Final Model Comparison

All the three models were retrained with the 4 chosen features and optimum hyperparameters.

7.1 Comparison Table

Model	Features	CV RMSE	Test MSE	Test RMSE	Test MAE	Test R^2
MLP Regressor	4	53.72	2818.72	53.09	32.05	0.4104
Random Forest (Tuned)	4	38.71	1355.51	36.82	20.67	0.7165
Linear Regression	4	64.71	4329.89	65.80	40.44	0.0943

7.2 Analysis

Best Model: Random Forest (Tuned) — $R^2 = 0.7165$

Key observations:

Random Forest was always the best in regards to all measures, reaching the lowest RMSE (36.82), lowest MAE (20.67), and highest R 2 (0.7165). The ensemble feel of the same actually nails the non-linear interactions and feature interactions of the climate data. The performance of MLP Regressor was moderate (R 2 = 0.4104) which was way below that of the Random Forest. This implies that neural nets can learn non-linear items, however, this particular MLP may require additional data, additional features, or greater fine-tuning to remain with ensemble techniques on this data. Linear Regression failed (R 2=0.0943), which shows that linear models are inadequate to this problem. The relations of the climate data are essentially non-linear and entangled relations between the location of the station, type of element, and time relations, which can not be described using linear models. The difference between the Random Forest (R 2 = 0.72) and Linear Regression (R 2 = 0.09) reveals the importance of critical model choice when dealing with complicated data. In this non-linear prediction problem on climatic analysis, over 7x more variance can be locked in. Even with 4 features only, the results were very near to all the 8 features - dimensionality reduction is completely achievable without key information.

8. Conclusion and Reflection

8.1 Summary of Findings

The project indicates how machine learning can be applied to SDG 13 Climate Action by training regression models to compute the values of climate measurements based on the GHCN historical dataset. There are three models developed and compared: MLP Regressor A three-layer neural network (1286432neurons) with ReLU and early stopping, achieving R 2 = 0.41. Random Forest Regressor - this is an ensemble of 200 trees, which are configured through GridSearchCV, which has the best output, and the R2 = 0.72. Linear Regression- a minimum model that indicates the constraints of linear methods (R 2 = 0.09)

8.2 Key Insights

More than 90 per cent of the predictive power is due to element type and station location which are the large drivers of climate value prediction. Also, non-linear models (Random Forest, MLP) eliminate the linear ones, which proves that the relationships between the climate data are complicated. • The top 4 discriminative features emerged during mutual information-based feature selection and allowed us to maintain the efficiency of the predictions with fewer and less computation. • Hyper-parameter tuning on the cross-validation enhanced the Random Forest and provided sound generalisation estimate

8.3 Limitations

• The 1901 is the only year the dataset is available, which is why we cannot claim the model is effective in the long-term. • The feature space is sparse - more meteorological variables would come in handy. • It could require additional tuning of the MLP, or alternative architecture (e.g.

LSTM over time). Various scales of measurement (temperature, precipitation) differ in values and may poison the model.

8.4 Future Work

• Fetch the data over more years to pick up longer climate trends. • test time-series models such as LSTM or GRU to more effectively capture time dependency. • Add spatial information of geographic coordinates and elevation data. • Develop model individual to each type of element to be more specialised. • Deep learning experiments on large scale climatic prediction. 8.5 SDG13- Climate Action Contribution. This project demonstrates the use of AI and machine learning to make sense of the historic climate data made to support SDG 13 by: • Driving insights on the patterns of climate and station-based variation. - Developing forecasting tools that can support the monitoring of climate and the early warning mechanisms. In general, stressing the strength of non-linear environmental relationship models. • Contributing to the increase in AI-based climate science.

9. References

1. Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E. and Houston, T.G. (2012) 'An overview of the Global Historical Climatology Network-Daily database', Journal of Atmospheric and Oceanic Technology, 29(7), pp. 897–910.
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J. (2011) 'Scikit-learn: Machine Learning in Python', Journal of Machine Learning Research, 12, pp. 2825–2830.
3. Breiman, L. (2001) 'Random forests', Machine Learning, 45(1), pp. 5–32.
4. Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. Cambridge, MA: MIT Press.
5. United Nations (2015) Transforming Our World: The 2030 Agenda for Sustainable Development. Available at: https://sdgs.un.org/goals/goal13 (Accessed: February 2026).
6. Rolnick, D., Donti, P.L., Kaack, L.H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A.S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A. and Luccioni, A. (2022) 'Tackling climate change with machine learning', ACM Computing Surveys, 55(2), pp. 1–96.

SDG 13 — Climate Action: "Take urgent action to combat climate change and its impacts."