Next-Day Average Temperature Prediction in Rwanda Using Machine Learning and Deep Learning Techniques

GAHIGI Robert
Student ID: 101101
MSc Big Data Analytics
Adventist University of Central Africa (AUCA)

Academic Year 2025–2026

Abstract

Accurate short-term temperature forecasting is essential for agriculture planning, disaster preparedness, and environmental monitoring in Rwanda. This study develops and compares multiple machine learning models to predict next-day average temperature using historical meteorological observations. Key predictors include **Temp_lag1, Temp_lag2, Precip(precipitation), RH(Relative Humidity), SUNSHINE, Cloud_opacity, Windspeed, and Vapour_pressure**. After rigorous data cleaning and median imputation, regression, classification, and clustering analyses were performed. The Multi-Layer Perceptron(MLP) Regressor achieved the best regression performance with RMSE = 1.205517 and R^2 = 0.736358. Random Forest showed strong classification performance (ROC-AUC = 0.920614), while K-Means clustering identified three meaningful weather regimes (silhouette = 0.2556).Furthermore Hyperparameter tuning on Random Forest as Model improvement technique enhanced the predictive accuracy. The finding confirm that recent temperature history(Temp-lag1) is the dominant predictor, also demonstrate that machine learning models can effectively capture short-term atmospheric dynamics and provide reliable temperature forecasts for Rwanda.

Keywords: Temperature forecasting, Machine learning, Rwanda weather, Regression, Clustering

1 Introduction

Weather variability significantly affects agriculture, water management, and climate risk planning in Rwanda. Reliable short-term temperature prediction supports evidence-based decision making in climate-sensitive sectors such as crop management and hydrological planning. Traditional statistical forecasting approaches often struggle to capture nonlinear atmospheric dynamics. With the increasing availability of meteorological data, machine learning provides powerful tools for learning complex temporal relationships.Machine learning methods such as linear models, tree-based methods, and neural networks have become standard tools for predictive modeling in modern statistical learning [3]. Previous studies have shown that ensemble and neural models can provide reliable short-term forecasts when temporal features are properly engineered [4, 5, 6, 7]..

2 OBJECTIVES

2.1 The primary objective of this study

is to develop and compare multiple machine learning models for next-day temperature prediction in Rwanda.

2.1.1 Specific objectives:

- Perform data cleaning and exploratory analysis
- Build regression and classification models
- Identify weather regimes(atmospheric conditions) using clustering
- Improve and compare model performance

3 Dataset description

The dataset contains more than 108,000 daily meteorological observations including temperature,precipitation ,relative Humidity ,sunshine duration , wind speed, cloud opacity, and vapor pressure collected from multiple stations across Rwanda.Meteorological measurements follow standards defined by the world Meteorological Organization[2]

3.1 Data Pre-processing

Data preparation involved:

- Removal of variables with excessive missing values

- Median imputation to reduce outlier influence
- Creation of lag features (Temp_lag1, Temp_lag2)
- Seasonal feature extraction (month, day of year)

Important note: Temperature values $\leq 0°C$ were treated as missing because such values are unrealistic under Rwanda climate conditions and likely represent sensor or recording errors.

Missing predictor values were imputed using the median of each variable because the median is robust to extreme values and skewed meteorological distributions.

3.2 Feature Engineering

To capture temporal persistence in atmospheric conditions, two lag features were created:

$$\text{Temp\_lag1} = T_{t-1}, \quad \text{Temp\_lag2} = T_{t-2}$$

The prediction target is:

$$T_{t+1} = \text{next-day average temperature}$$

These lag variables allow the models to learn short-term memory effects that are common in weather time series.

4 Exploratory Data Analysis

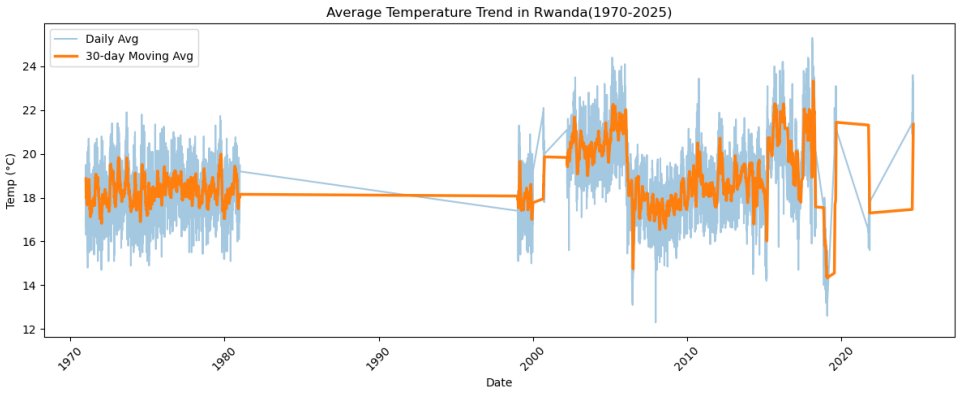

Figure 1: Average daily temperature since 1970–2025.

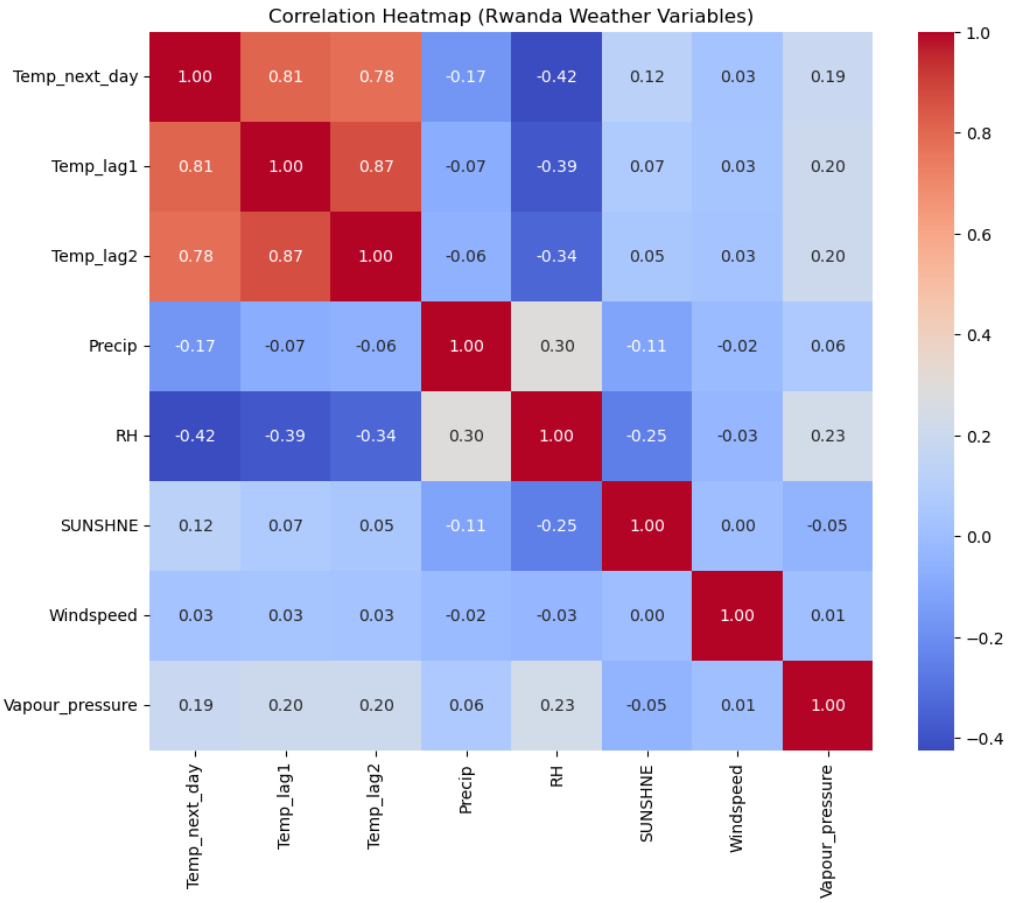

Figure 2: Correlation matrix of meteorological predictors.

The correlation analysis reveals that **Temp_lag1** and **Temp_lag2** have the strongest positive relationship with the next-day temperature. This indicates strong temporal persistence, meaning that recent temperature values are highly informative for short-term forecasting. Other atmospheric variables such as relative humidity and vapour pressure show moderate correlations, suggesting secondary influence on temperature dynamics.

5 Methodology

5.1 Models

The study evaluates both linear and nonlinear machine learning models:

- Linear Regression
- Random Forest Regressor
- MLP Regressor
- Logistic Regression (hot-day classification)
- K-Means clustering (weather regimes)

5.2 Evaluation Metrics

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum (y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

These evaluation metrics are standard measures for regression model assessment in statistical learning and time-series forecasting [7, 8, 1].

6 Results and Interpretation

6.1 Regression Performance

Table 1: Regression Results

Model	MAE	RMSE	R^2
MLP Regressor	0.911925	1.205517	0.736358
Linear Regression	0.952955	1.250719	0.716216
Random Forest(before tuning)	0.940654	1.258784	0.712545
Random Forest (After Tuning)	0.911084	1.218358	0.730711

The results from the table above shows that Multi-Layer Perceptron(MLP) Regressor(deep learning) achieved the lowest RMSE, indicating the smallest average prediction error among the evaluated models. Its R^2 value of 0.736358 means that approximately 73.6% of the variability in next-day temperature is explained by the model.

The Random Forest model also performed competitively after hyperparameter tuning, suggesting that nonlinear relationships exist in the data. Linear Regression, although slightly weaker, still achieved reasonable performance due to the strong autocorrelation structure in temperature time series.

6.2 Model Improvement

6.2.1 Hyperparameter tuning on Random Forest performance

To further improve model performance, hyperparameter tuning was applied to the Random Forest Regressor. Tuning aims to identify the optimal combination of model settings that minimizes prediction error while maintaining good generalization.

After tuning, the Random Forest model showed clear improvement, with RMSE decreasing from 1.258784 to 1.218358 and R^2 increasing from 0.712545 to 0.730711. This confirms that proper configuration of ensemble models enhances predictive accuracy.

Table 2: Random Forest hyperparameters selected after tuning.

Hyperparameter	Value
$n\_estimators$	500
max_depth	50
$min\_samples\_split$	2
$min\_samples\_leaf$	4
$max\_features$	auto
$bootstrap$	True
$random\_state$	42

Table 3: Impact of hyperparameter tuning on Random Forest performance.

Model Version	MAE	RMSE	R^2
Random Forest (Before Tuning)	0.940654	1.258784	0.712545
Random Forest (After Tuning)	0.911084	1.218358	0.730711

6.3 Classification Performance

Table 4: Hot-Day Classification

Model	Accuracy	F1	ROC-AUC
Random Forest	0.848864	0.866911	0.920614
MLP Classifier	0.845833	0.864198	0.921607
Logistic Regression	0.832389	0.849464	0.916512

The Random Forest classifier achieved the highest accuracy, indicating strong capability in distinguishing hot and non-hot days.

6.4 Clustering Analysis

$$S = 0.2559$$

This silhouette score indicates moderate but meaningful cluster separation.

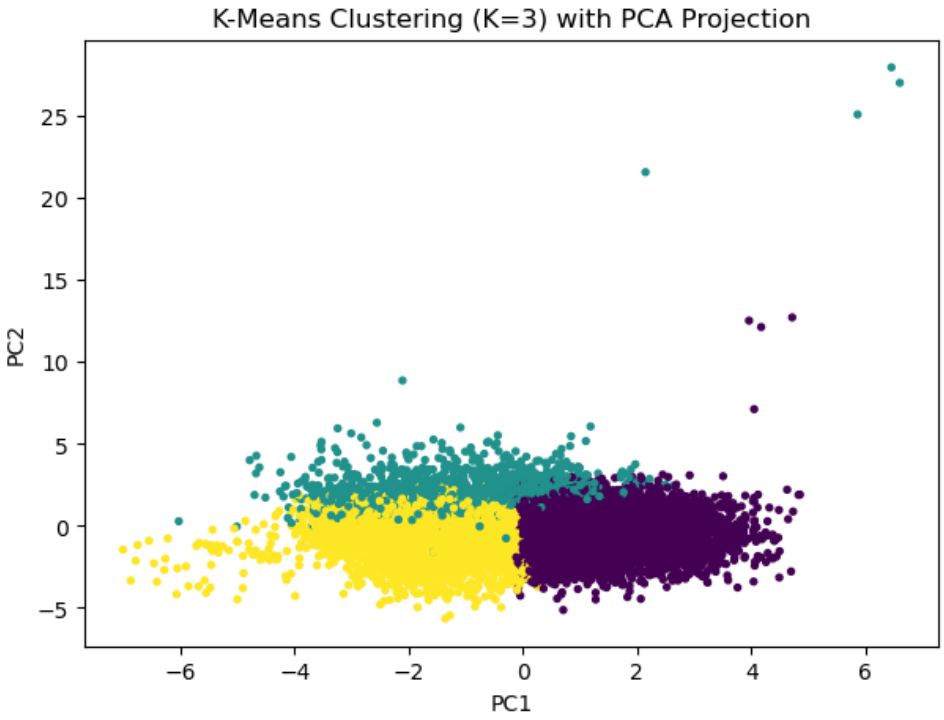

Figure 3: These clusters likely represent dry, moderate, and wet atmospheric onditions.

6.5 Feature Importance

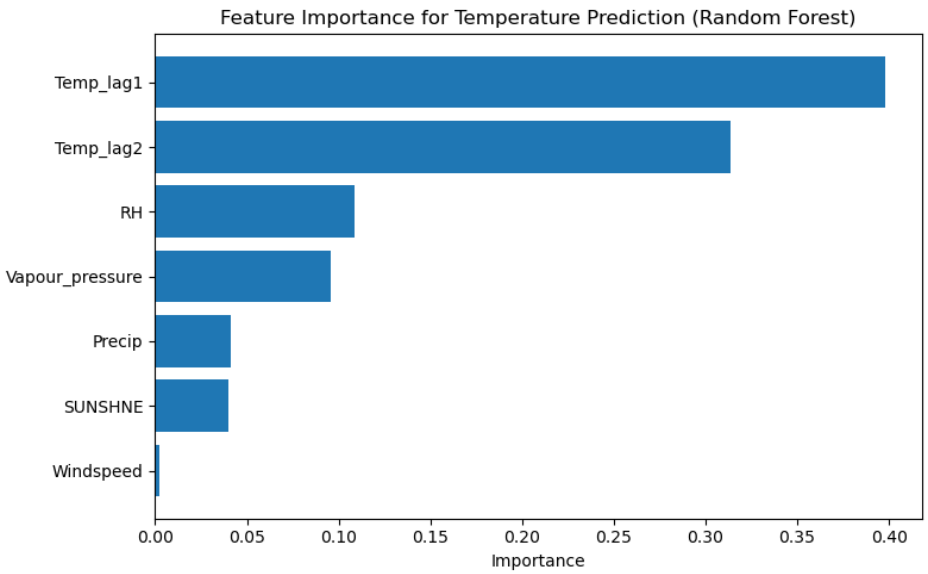

Figure 4: Feature importance from Random Forest.

Feature importance from Random Forest model shows that Temp_lag1 and Temp_lag2 clearly dominate the prediction process, confirming strong temporal persistence in Rwanda temperature dynamics.

==========

7 Discussion and Recommendations

This study confirms that nonlinear machine learning models significantly improve temperature forecasting accuracy in Rwanda.

Key findings:

- Neural networks achieved the best regression accuracy.
- Random Forest performed best for classification.
- Temperature persistence is the dominant predictive signal.
- Weather regimes(weather conditions) show moderate separability.

Recommendations:

- Deploy ensemble or neural models operationally.
- Incorporate additional atmospheric predictors.
- Extend the framework to rainfall forecasting.
- Investigate LSTM models for temporal deep learning.

8 Conclusion

Machine learning provides an effective framework for short-term temperature prediction in Rwanda. The comparative analysis demonstrates the superiority of nonlinear models and highlights the importance of temporal feature engineering for meteorological forecasting.

References

[1] Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python.

[2] World Meteorological Organization (2019). Guide to Meteorological Instruments.

[3] James, Gareth, Witten, D., Hastie, T., Tibshirani, Robert and Taylor, and Taylor, Jonathan (2023). *An Introduction to Statistical Learning with Applications in Python*. Springer, Cham.

[4] Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences* (3rd ed.). Academic Press.

[5] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

[6] Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks for modelling and prediction: A review of applications to environmental sciences. *Atmospheric Environment*, 32(14–15), 2627–2636.

[7] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer.

[8] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.