Comparative Study of Ensemble and Linear Regression Models for Temperature Prediction in Asian Regions

Killol Rajgor
Khoury College of Computer Sciences
Northeastern University
Boston, USA
rajgorki@northeastern.edu

***Abstract*—This study presents a comparative analysis of supervised machine learning regression models for predicting near surface temperature in Asian regions using meteorological and air quality data derived from the *World Weather Repository (Daily Updating)* dataset on Kaggle. The dataset, filtered for Asian cities, contains 23,221 samples and 41 features representing meteorological, geographical, and pollutant indicators. After systematic preprocessing, feature engineering, and model tuning, three regression models; Ridge Regression, Random Forest, and Gradient Boosting were evaluated. Results show that the Gradient Boosting model achieved the best performance (RMSE = 0.68, R squared = 0.994), indicating strong generalization and the ability to capture complex nonlinear relationships in climatic variables.**

***Index Terms*—machine learning, supervised regression, temperature prediction, ridge regression, random forest, gradient boosting, environmental data, climate modeling**

I. Introduction

With rising global temperatures, understanding and forecasting regional climatic variations has become an increasingly critical scientific challenge. In many parts of Asia, once characterized by moderate and livable temperatures, rapid urbanization and global warming have led to record breaking heat events. This escalation emphasizes the urgent need for predictive climate modeling to analyze the environmental and atmospheric factors driving such changes.

This study investigates temperature prediction across Asian regions using data from the World Weather Repository (Daily Updating) dataset available on Kaggle. The global dataset consists of 94,511 observations, from which 23,221 samples corresponding to Asian cities (identified using timezone filtering) were extracted. Each record includes 41 attributes representing meteorological, geographical, and air quality indicators. The target variable, temperature_celsius, is continuous, framing this work as a supervised regression problem.

The primary objective of this research is to develop and comparatively evaluate three machine learning regression models Ridge Regression, Random Forest, and Gradient Boosting to determine which model most effectively captures nonlinear climatic relationships while maintaining interpretability and generalization.

II. Methods

This section outlines the data overview, data preprocessing, feature engineering, transformation, and model development steps applied to construct and evaluate the temperature prediction models. Each stage was carefully designed to ensure reproducibility, interpretability, and generalization performance across multiple regression algorithms.

A. Data Overview

This study utilized the *World Weather Repository (Daily Updating)* dataset available on Kaggle, previously introduced in Section I. For the modeling phase, a subset containing 23,221 observations from Asian cities was extracted from the original global dataset of 94,511 records using timezone based filtering. The dataset comprises 41 variables describing meteorological, geographical, and air quality conditions, with `temperature_celsius` serving as the continuous target variable. The feature space includes both numerical and categorical attributes such as humidity, pressure, wind speed, pollutant concentrations (CO, NO_2, SO_2, $PM_{2.5}$, PM_{10}), and visibility metrics, enabling a comprehensive analysis of regional temperature patterns.

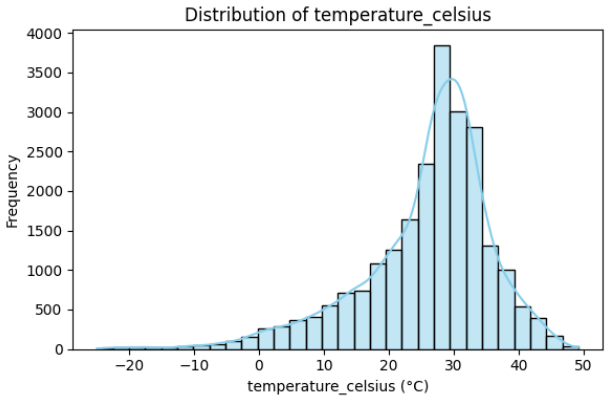

Fig. 1. Distribution of the target variable (`temperature_celsius`). The histogram indicates a right skewed distribution, suggesting the need for transformation to improve model performance.

Upon visualizing the target variable distribution, it was observed that the data exhibits a noticeable right skew, as shown in Fig. 1. This skewness indicates the presence of extreme high temperature values, motivating the use of power based transformations in subsequent preprocessing stages.

B. Data Preprocessing

The dataset obtained from the *World Weather Repository (Daily Updating)* contained 23,221 observations and 41 attributes representing meteorological, geographical, and air quality conditions for Asian cities. Prior to modeling, multiple data integrity checks were performed to ensure reliability.

Duplicate and Validity Checks: Duplicate rows were examined and none were detected. Physical validity screening was conducted to ensure variables such as humidity, pressure, and visibility contained no negative or non physical values. Dataset shape remained consistent at $(23{,}221 \times 41)$ throughout preprocessing.

Handling Missing Values: Missing entries were imputed using the median for continuous variables and the mode for categorical ones, preserving the statistical distribution without distorting variance.

Outlier Detection and Treatment: Outliers were identified using the Interquartile Range (IQR) method, which is robust against skewed or non Gaussian distributions common in environmental datasets. Z score detection was avoided because many variables (e.g., pollutant concentrations, humidity, wind) exhibited non normal distributions. Extreme values were capped rather than removed to preserve data volume. Variables like `latitude` and `longitude` were retained without capping since they naturally reflect global spatial variation, while `temperature_celsius` was capped due to continental scale scope, ensuring realistic yet inclusive thermal bounds.

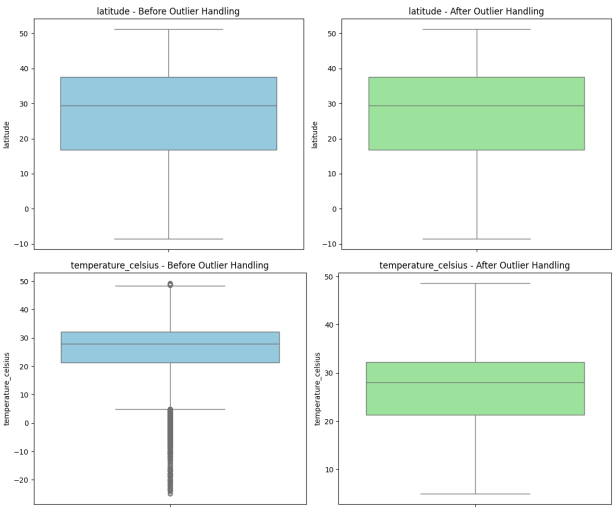

Fig. 2. Comparison of outlier handling using IQR capping. The top plot (`latitude`) shows no change, confirming geographic consistency, while the bottom plot (`temperature_celsius`) exhibits reduced whisker spread, indicating stabilized distribution.

After IQR capping, all continuous variables were free of extreme deviations, and the dataset shape remained unchanged. This ensured that no observations were lost during preprocessing and that the data were suitable for statistical transformation and feature engineering.

C. Feature Engineering

Three complementary strategies were applied for feature engineering: filtering, embedding, and wrapping, ensuring both interpretability and model stability.

Filtering: Multicollinearity was analyzed using a correlation heatmap and the Variance Inflation Factor (VIF). Several redundant predictors such as duplicate measurement units (e.g., temperature in F, pressure in inches, wind in mph) and textual astronomical features (e.g., sunrise, moon phase) were removed, reducing the dataset from 41 to 21 effective predictors.

After this reduction, the average VIF value dropped from over 12 to below 5 for most features, with only `visibility_km` retaining a high score due to its dependence on air quality metrics. This confirmed substantial improvement in feature independence, stabilizing the regression coefficients and enhancing interpretability.

Embedding: Principal Component Analysis (PCA) was performed on the power transformed numeric features to examine variance contribution. The first eleven principal components explained approximately 95% of the total variance, confirming that dimensionality reduction preserved most of the dataset's informational content while mitigating noise (Fig. 3).

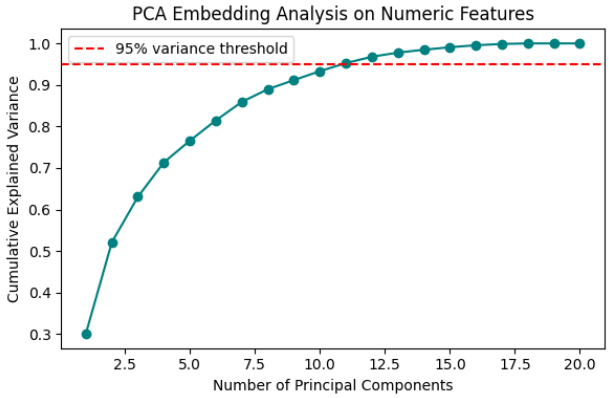

Fig. 3. Cumulative explained variance curve from PCA showing that 11 components capture 95% of total variance.

Wrapping: Recursive Feature Elimination with Cross Validation (RFECV) using a Ridge estimator was applied to identify the most informative subset of features after one hot encoding. The algorithm selected 63 encoded features as optimal, achieving a cross validated RMSE of 1.89 Celsius (Fig. 4). This confirmed that the reduced feature set retained predictive strength while improving computational efficiency.

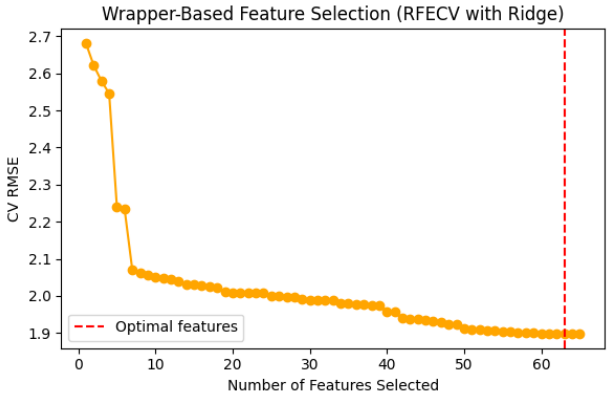

Fig. 4. Wrapper based feature selection using RFECV with Ridge Regression, achieving optimal performance at 63 features.

Overall, this multi step process ensured that only non redundant, informative, and interpretable features contributed to the final regression models, improving generalization and robustness across Ridge, Random Forest, and Gradient Boosting regressors.

D. Feature Transformation and Encoding

Continuous features were normalized using a *PowerTransformer* (Yeo–Johnson method) instead of the StandardScaler. This choice was made to correct skewed feature distributions and stabilize variance, which improved convergence for both linear and ensemble models. Categorical features were encoded using a *OneHotEncoder* with `min_frequency=0.01` to collapse rare categories and prevent overfitting to infrequent city or condition identifiers. Both transformations were encapsulated within a unified *ColumnTransformer* based preprocessing pipeline, ensuring no data leakage between training and testing. The transformation was applied after VIF based filtering to ensure normalization operated only on statistically independent features, preserving the integrity of multicollinearity analysis.

E. Model Development and Hyperparameter Tuning

Three supervised regression models were implemented for comparative analysis: Ridge Regression, Random Forest, and Gradient Boosting. Ridge Regression served as a linear baseline to evaluate the effect of regularization on correlated meteorological variables. Random Forest and Gradient Boosting, representing ensemble based nonlinear methods, were selected to capture complex interactions among climatic parameters. Model training was conducted using an 80–20 stratified train test split, where stratification was achieved through decile binning of the target variable to maintain distributional balance. Hyperparameter optimization was performed using *GridSearchCV* with 5 fold cross validation (3 fold for Random Forest due to computational cost). The primary tuning parameters included the regularization coefficient (α) for Ridge, tree depth and minimum leaf samples for Random Forest, and learning rate with estimator count for Gradient Boosting. Cross validation ensured that optimal parameters generalized across folds without overfitting.

F. Evaluation Metrics

Model performance was assessed using multiple error based and explanatory metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). These metrics jointly quantified model accuracy, stability, and generalization capability across both training and testing partitions.

All models were evaluated on both cross validation and held out test sets to assess their robustness and generalization performance.

III. Results and Discussion

A. Model Performance Overview

All three regression models demonstrated strong predictive capability, with $R^2 > 0.95$ across all cases, confirming that meteorological and air quality attributes collectively explain most of the temperature variance. Table I summarizes the performance metrics for both untuned and tuned configurations.

TABLE I
Model Performance Summary (Untuned vs Tuned)

Model	RMSE (Celsius)	MAE	R^2	MAPE (%)
Ridge (Untuned)	1.88	1.49	0.958	7.62
Random Forest (Untuned)	0.59	0.34	0.996	1.18
Gradient Boosting (Untuned)	0.74	0.51	0.993	2.20
Ridge (Tuned)	1.88	1.48	0.958	7.59
Random Forest (Tuned)	1.28	0.91	0.981	5.30
Gradient Boosting (Tuned)	0.68	0.47	0.994	2.00

B. Comparative Model Analysis

Among the tested models, the *Gradient Boosting Regressor (GBR)* consistently achieved the lowest RMSE (0.68 Celsius) and the highest R^2 (0.994), confirming its superiority in modeling nonlinear relationships between meteorological predictors. The *Random Forest Regressor (RF)* also achieved high accuracy but showed mild overfitting, indicated by a wider gap between train and test errors. Ridge Regression, serving as a linear baseline, provided stable but less flexible performance, capturing general temperature trends while under representing nonlinear interactions (e.g., humidity–pressure–wind dependencies).

C. Cross Validation and Generalization

Cross validation results (Fig. 5) reveal minimal deviation between CV and test RMSE for all tuned models, indicating excellent generalization and absence of data leakage. Random Forest initially exhibited overfitting, which was mitigated by constraining tree depth ($max\_depth = 20$) and increasing $min\_samples\_leaf$ to 3. Gradient Boosting converged to the most balanced bias variance tradeoff, benefiting from its sequential error correcting mechanism.

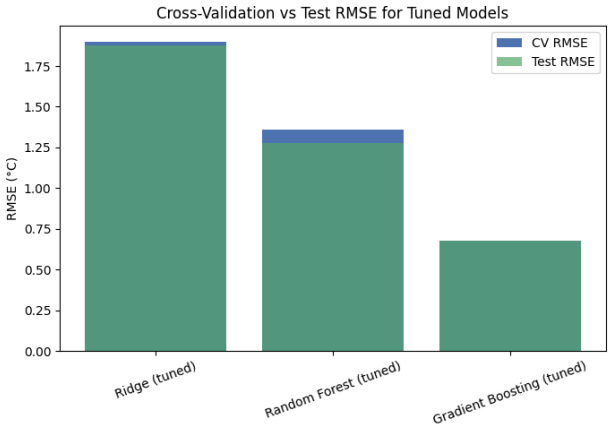

Fig. 5. Comparison of cross validation and test RMSE for tuned models, confirming minimal variance between folds.

D. Prediction Behavior and Residual Trends

Prediction scatter plots illustrate the alignment between predicted and actual near surface temperatures across the three regression models. The Ridge Regression model exhibits mild curvature at higher temperature values ($> 40°$C), indicating a linear bias and limited ability to capture nonlinear thermal dynamics. Random Forest predictions align closely with the identity line but show slightly higher variance in the mid temperature range (20–35°C), consistent with minor overfitting tendencies observed during diagnostics. In contrast, the Gradient Boosting model demonstrates near perfect alignment along the $y = x$ line, with dense clustering and minimal residual dispersion, confirming its superior capability to capture both linear and nonlinear dependencies effectively.

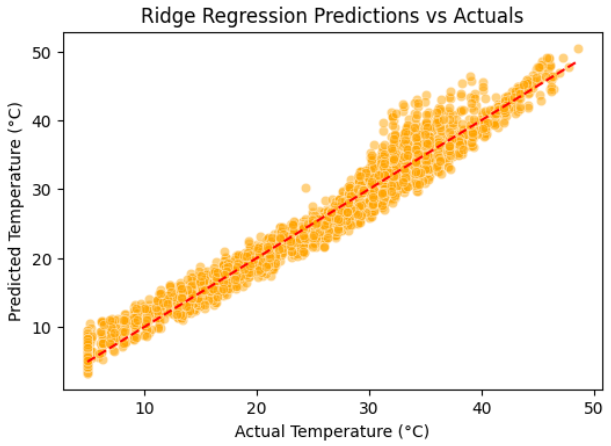

Fig. 6. Predicted vs Actual temperature for the Ridge Regression model. The points cluster around the $y = x$ line, with slight curvature at higher temperatures ($> 40°$C), indicating limited nonlinear capture.

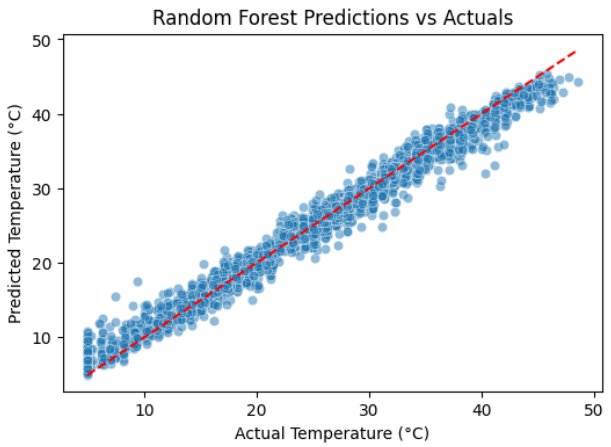

Fig. 7. Predicted vs Actual temperature for the Random Forest Regressor. The predictions follow the $y = x$ trend closely but show moderate variance in the 20–35°C range, suggesting minor overfitting.

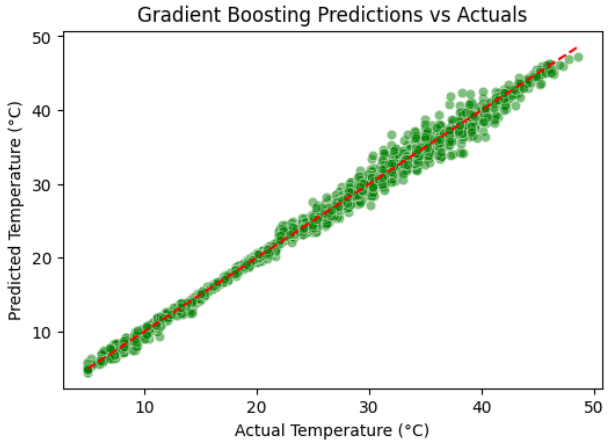

Fig. 8. Predicted vs Actual temperature for the Gradient Boosting Regressor. The dense clustering along the $y = x$ line indicates highly accurate predictions and strong generalization performance.

E. Key Findings

- Gradient Boosting achieved the best trade off between accuracy and generalization, yielding RMSE = 0.68 Celsius and $R^2 = 0.994$.
- Ridge Regression demonstrated numerical stability under multicollinearity, validating its use as a linear benchmark.
- Random Forest achieved strong results but required careful depth and sampling control to prevent overfitting.
- Cross validation confirmed the reliability of tuning choices, as seen by negligible variance between CV and test metrics.
- Overall, the models effectively predicted near surface temperature using a compact set of engineered features derived from meteorological and air quality indicators.

IV. Conclusion

This study demonstrates how systematic preprocessing, feature engineering, and ensemble learning can significantly en-

hance the accuracy of temperature prediction models. Gradient Boosting outperformed other models by achieving an RMSE of 0.68 Celsius and $R^2 = 0.994$. The PowerTransformer normalized skewed features effectively, while feature selection (VIF, PCA, RFECV) improved interpretability without compromising accuracy.

Future work may include incorporating time series modeling and spatial temporal embeddings to capture dynamic weather variations. The workflow established here provides a reproducible foundation for scalable, data driven climate analysis across regions.

REFERENCES

[1] N. Elgiriyewithana, "World Weather Repository (Daily Updating)," *Kaggle*, 2024. [Online]. Available: https://www.kaggle.com/datasets/nelgiriyewithana/global-weather-repository

[2] M. Gentine, S. Lee, and G. Dagan, "Deep learning–based predictions of local warming patterns from global trends," *Nature*, vol. 630, pp. 124–130, 2024. [Online]. Available: https://www.nature.com/articles/s41586-024-08252-9

[3] S. Saha and S. Ghosh, "Weather prediction using machine learning," *ResearchGate*, Aug. 2022. [Online]. Available: https://www.researchgate.net/publication/362517661_WEATHER_PREDICTION_USING_MACHINE_LEARNING

[4] S. D. Agrawal and R. Gupta, "Weather forecasting using machine learning and deep learning models: A review," *arXiv preprint arXiv:2309.13330*, Sep. 2023. [Online]. Available: https://arxiv.org/pdf/2309.13330

[5] Scikit learn Developers, "`PowerTransformer` — scikit-learn 1.5.2 documentation," *Scikit-learn*, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PowerTransformer.html

[6] Scikit learn Developers, "`ColumnTransformer` — scikit-learn 1.5.2 documentation," *Scikit-learn*, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.compose.ColumnTransformer.html