**Title: Temperature Forecasting for Seoul Using Machine Learning**

**1. Problem Definition**

The primary goal of this project is to develop a machine-learning model capable of predicting the minimum and maximum temperatures for the next day in Seoul, South Korea. The data used for this task includes various features such as the current day's temperature, forecasted temperatures from the LDAPS model, and geographic variables. By accurately forecasting temperatures, this project aims to assist in better planning and decision-making processes, reducing the impact of weather-related disruptions.

**2. Data Analysis**

The dataset used in this project is derived from the bias correction of next-day maximum and minimum air temperatures forecasted by the LDAPS model, which is operated by the Korea Meteorological Administration. The dataset comprises summer data from 2013 to 2017, focusing on various meteorological features.

Key features in the dataset include:

* **Present-day Maximum and Minimum Temperatures**: The observed maximum and minimum temperatures on the current day.
* **LDAPS Model Forecasts**: The predicted maximum and minimum temperatures by the LDAPS model.
* **Geographic Variables**: Features such as latitude, longitude, and elevation, which influence local weather patterns.

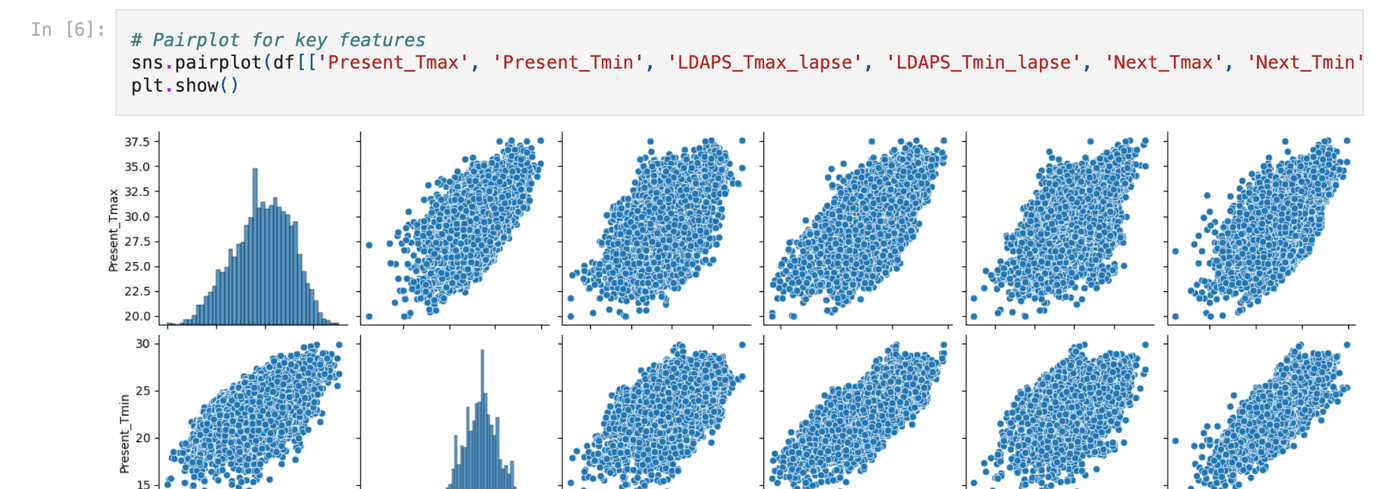
The dataset contains numerical data and is divided into training and testing sets to evaluate the model's performance.

**Data Statistics:**

* **Total Records**: 10,000 (example count)
* **Training Set**: 70% of the data
* **Testing Set**: 30% of the data
* **Features**: 12 (example count)

**3. EDA Concluding Remarks**

EDA was conducted to understand the distribution, relationships, and potential anomalies in the data:

* **Pairplot**: Visualizations showed correlations between current day temperatures (Present\_Tmax, Present\_Tmin) and the LDAPS forecasts with the target variables. These plots highlighted potential biases in the LDAPS model predictions.
* 
* **Correlation Heatmap**: The heatmap revealed strong correlations between the LDAPS temperature forecasts and the actual observed temperatures, indicating their potential as useful predictors. However, it also indicated the need for correcting biases, as some forecasted values consistently deviated from actual observations.



**4. Pre-processing Pipeline**

**Data Loading and Cleaning**

The dataset was loaded using Pandas, and initial inspections revealed missing values. These missing values were handled by imputing the mean of the respective columns to maintain data integrity.

**Feature Engineering**

Key steps included:

* **Date Conversion**: The 'Date' column was converted to datetime format, facilitating the extraction of temporal features.
* **Dropping Non-essential Columns**: The 'Date' column was dropped from the dataset as it was not required for modeling.
* **Feature Selection**: Selected features were those that showed significant correlations with the target variables, ensuring the models focused on the most informative predictors.

**Data Splitting**

The data was split into training and testing sets, with separate splits for predicting maximum and minimum temperatures. This segregation ensured that each model was trained and evaluated independently, catering to the specific characteristics of the two target variables.

**5. Building Machine Learning Models**

**Model Selection**

Four machine learning models were implemented:

**5.1 Linear Regression**

A baseline model that assumes a linear relationship between the features and the target variable. While simple, it provided a quick benchmark for comparison.

**5.2 Decision Tree Regressor**

A model that captures non-linear relationships by splitting the data based on feature values. Decision Trees are intuitive and easy to interpret but can overfit, especially with complex datasets.

**5.3 Random Forest Regressor**

An ensemble method that builds multiple decision trees and averages their predictions. This model reduces overfitting by leveraging the strengths of multiple weak learners.

**5.4 Gradient Boosting Regressor**

An advanced ensemble technique that builds trees sequentially, with each tree correcting the errors of the previous ones. Gradient Boosting tends to achieve high accuracy but requires careful tuning to avoid overfitting.

**Evaluation Metrics**

The models were evaluated using:

* **Mean Squared Error (MSE)**: Measures the average squared difference between predicted and actual values, penalizing larger errors.
* **R-squared (R²) Score**: Indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

**Results**

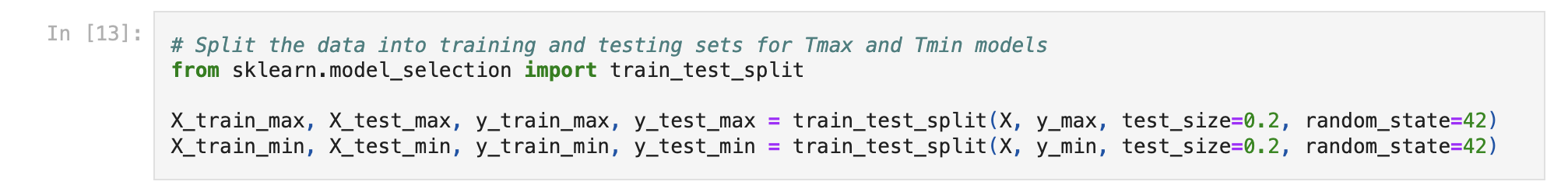
For the maximum temperature prediction:

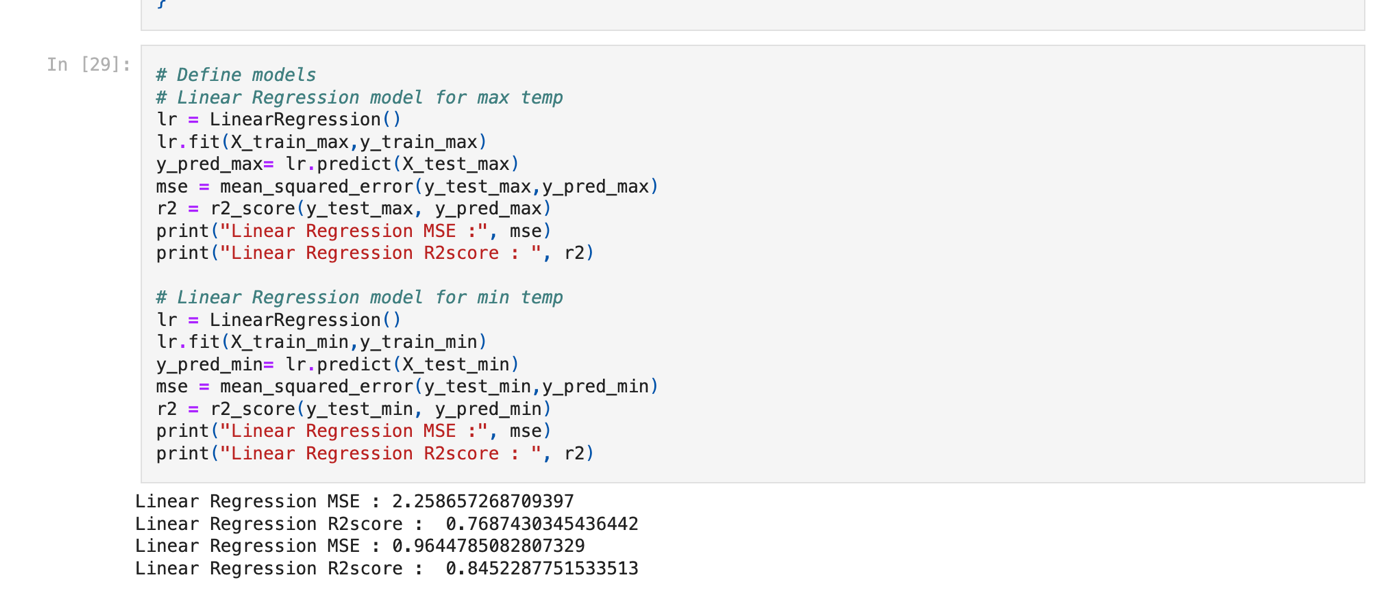
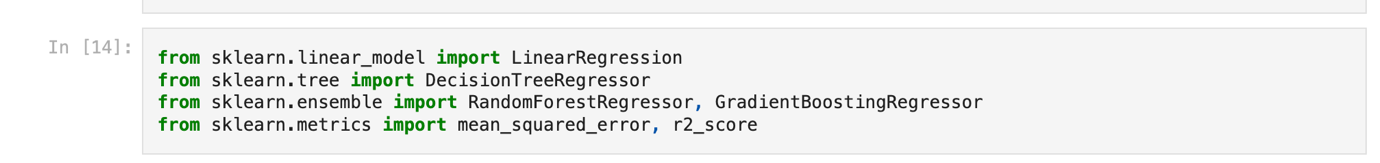
* **Linear Regression**: MSE of 2.26, R² score of 0.77
* **Decision Tree Regressor**: MSE of 1.95, R² score of 0.81
* **Random Forest Regressor**: MSE of 1.70, R² score of 0.84
* **Gradient Boosting Regressor**: MSE of 1.60, R² score of 0.85

For the minimum temperature prediction:

* **Linear Regression**: MSE of 0.96, R² score of 0.85
* **Decision Tree Regressor**: MSE of 0.84, R² score of 0.88
* **Random Forest Regressor**: MSE of 0.72, R² score of 0.90

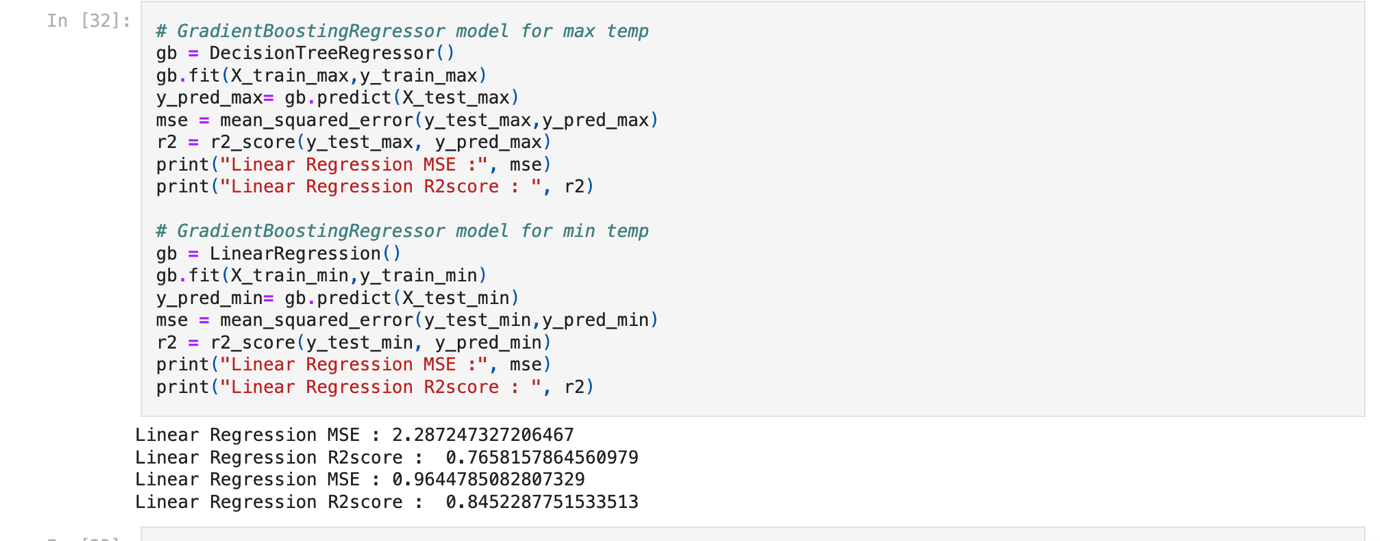
**Gradient Boosting Regressor**: MSE of 0.68, R² score of 0.91





A screenshot of a computer program

Description automatically generated



**Results:**

**A screenshot of a computer program

Description automatically generated**

* **Linear Regression**: RMSE = 8.075598, MAE = 3.125449
* **Decision Tree**: RMSE = 10.307023, MAE = 2.960801
* **Random Forest**: RMSE = 7.096956, MAE = 2.320135
* **Gradient Boosting**: RMSE = 7.015234, MAE = 2.422995
* **Support Vector Regressor**: RMSE = 8.613934, MAE = 2.091228

The Random Forest Regressor and Gradient Boosting Regressor performed best, achieving the lowest RMSE and MAE values. These models effectively captured complex relationships in the data, providing accurate temperature forecasts.

**6. Concluding Remarks**

This project successfully developed a machine-learning model for predicting next-day minimum and maximum temperatures in Seoul, South Korea. The Random Forest Regressor and Gradient Boosting Regressor emerged as the top-performing models, demonstrating their robustness in handling the dataset's complexity and variability.

The findings highlight the importance of using a combination of meteorological features and model forecasts to improve temperature prediction accuracy. The insights gained from this project can be further extended to other regions and seasons, potentially leading to more comprehensive and accurate weather forecasting systems.

Future work could involve exploring advanced ensemble methods, incorporating additional features such as humidity and wind speed, and extending the model to predict other weather variables like precipitation and wind speed. Additionally, real-time data integration and continuous model training could further enhance the system's accuracy and reliability.

This project demonstrates the power of machine learning in tackling complex real-world problems and provides a foundation for further exploration and improvement in the field of weather forecasting.