***Article on Evaluation Project***

**Project Name: Temperature Forecast Project using ML**

**Project Description:**

This data is for the purpose of bias correction of next-day maximum and minimum air temperatures forecast of the LDAPS model operated by the Korea Meteorological Administration over Seoul, South Korea. This data consists of summer data from 2013 to 2017. The input data is largely composed of the LDAPS model's next-day forecast data, in-situ maximum and minimum temperatures of present-day, and geographic auxiliary variables. There are two outputs (i.e. next-day maximum and minimum air temperatures) in this data. Hindcast validation was conducted for the period from 2015 to 2017.

**Attribute Information:**

1. station - used weather station number: 1 to 25

2. Date - Present day: yyyy-mm-dd ('2013-06-30' to '2017-08-30')

3. Present\_Tmax - Maximum air temperature between 0 and 21 h on the present day (Â°C): 20 to 37.6

4. Present\_Tmin - Minimum air temperature between 0 and 21 h on the present day (Â°C): 11.3 to 29.9

5. LDAPS\_RHmin - LDAPS model forecast of next-day minimum relative humidity (%): 19.8 to 98.5

6. LDAPS\_RHmax - LDAPS model forecast of next-day maximum relative humidity (%): 58.9 to 100

7. LDAPS\_Tmax\_lapse - LDAPS model forecast of next-day maximum air temperature applied lapse rate (Â°C): 17.6 to 38.5

8. LDAPS\_Tmin\_lapse - LDAPS model forecast of next-day minimum air temperature applied lapse rate (Â°C): 14.3 to 29.6

9. LDAPS\_WS - LDAPS model forecast of next-day average wind speed (m/s): 2.9 to 21.9

10. LDAPS\_LH - LDAPS model forecast of next-day average latent heat flux (W/m2): -13.6 to 213.4

11. LDAPS\_CC1 - LDAPS model forecast of next-day 1st 6-hour split average cloud cover (0-5 h) (%): 0 to 0.97

12. LDAPS\_CC2 - LDAPS model forecast of next-day 2nd 6-hour split average cloud cover (6-11 h) (%): 0 to 0.97

13. LDAPS\_CC3 - LDAPS model forecast of next-day 3rd 6-hour split average cloud cover (12-17 h) (%): 0 to 0.98

14. LDAPS\_CC4 - LDAPS model forecast of next-day 4th 6-hour split average cloud cover (18-23 h) (%): 0 to 0.97

15. LDAPS\_PPT1 - LDAPS model forecast of next-day 1st 6-hour split average precipitation (0-5 h) (%): 0 to 23.7

16. LDAPS\_PPT2 - LDAPS model forecast of next-day 2nd 6-hour split average precipitation (6-11 h) (%): 0 to 21.6

17. LDAPS\_PPT3 - LDAPS model forecast of next-day 3rd 6-hour split average precipitation (12-17 h) (%): 0 to 15.8

18. LDAPS\_PPT4 - LDAPS model forecast of next-day 4th 6-hour split average precipitation (18-23 h) (%): 0 to 16.7

19. lat - Latitude (Â°): 37.456 to 37.645

20. lon - Longitude (Â°): 126.826 to 127.135

21. DEM - Elevation (m): 12.4 to 212.3

22. Slope - Slope (Â°): 0.1 to 5.2

23. Solar radiation - Daily incoming solar radiation (wh/m2): 4329.5 to 5992.9

24. Next\_Tmax - The next-day maximum air temperature (Â°C): 17.4 to 38.9

25. Next\_Tmin - The next-day minimum air temperature (Â°C): 11.3 to 29.8T

We had built separate models that can predict the minimum temperature for the next day and the maximum temperature for the next day based on the details provided in the dataset.

1. **Problem Definition:**

Accurate temperature forecasting is crucial for various sectors, including agriculture, energy management and public safety. Traditional methods rely on physical models of the atmosphere, which can be complex and computationally expensive. With the advent of machine learning, we can leverage historical weather data to predict future temperatures, potentially improving accuracy and efficiency. This project aims to develop a machine learning model to forecast temperature based on historical weather data. This data is for the purpose of bias correction of next-day maximum and minimum air temperatures forecast of the LDAPS model operated by the Korea Meteorological Administration over Seoul, South Korea. This data consists of summer data from 2013 to 2017. The input data is largely composed of the LDAPS model's next-day forecast data, in-situ maximum and minimum temperatures of present-day, and geographic auxiliary variables. This dataset helps us to predict the maximum temperature and minimum temperature for the next day.

1. **Data Analysis:**

The dataset contains 7752 rows and 27 columns.

The data analysis can be done in 3 methods:

1. **Univariate analysis**

By Plotting histogram for present\_Tmax and present\_Tmin variables we got to know about the fact that the minimum temperature is in the range (14-30) maximum temperature is in the range (21-38).

We also plotted a histogram for solar radiation and got to know that the range of solar radiation is (4250 - 6000).

We plotted histogram for NEXT\_Tmax and NEXT\_Tmin variables and we got the Next Day Maximum Temperature is in range of ('19 - 38'). Next Day Minimum Temperature is in range of ('13 - 30')

Count of average temperature is same in all the years.

1. **Bivariate analysis**

We observe variations between present maximum (Tmax) and minimum (Tmin) temperatures.

Yearly variations in maximum and minimum temperatures are not evident.

1. **Statistical Analysis:**

Statistical measures suggest the presence of outliers, indicated by discrepancies between mean and 50th percentile values. We can also check the skewness of the data through statistical analysis.

1. **Correlation Between the Variables:**

Present\_Tmax and Present\_Tmin exhibit a strong correlation with the target variables.

LDAPS\_Tmax\_lapse and LDAPS\_Tmin\_lapse also show significant correlations with the target variables.

LDAPS cloud cover and precipitation variables demonstrate negative correlations with the target variables.

1. **Outlier Detection:**

'Present\_Tmax', 'Present\_Tmin', 'LDAPS\_RHmin', 'LDAPS\_RHmax', 'LDAPS\_Tmax\_lapse', 'LDAPS\_Tmin\_lapse', 'LDAPS\_WS', 'LDAPS\_LH','LDAPS\_PPT1', 'LDAPS\_PPT2', 'LDAPS\_PPT3','LDAPS\_PPT4','DEM', 'Slope', 'Solar radiation', 'Month' have outliers present and can be removed using Z-score method.

1. **Skewness of the dataset:**

Present\_Tmax, Present\_Tmin, LDAPS\_RHmax, LDAPS\_Tmax\_lapse, and LDAPS\_Tmin\_lapse exhibit left skewness.

LDAPS\_WS, LDAPS\_LH, and other LDAPS variables related to cloud cover and precipitation display right skewness.

1. **EDA Concluding remarks:**

Exploratory Data Analysis (EDA) is an essential step in understanding the underlying patterns and relationships within a dataset. In this article, we perform EDA on a temperature forecast dataset, focusing on maximum and minimum temperatures, along with various meteorological variables.

Through univariate analysis, bivariate analysis, statistical analysis, and correlation analysis, we aim to gain insights into the dataset's characteristics and identify potential trends or patterns.

1. **Pre-processing pipeline:**

Data pre-processing is crucial for building an effective machine learning model. It involves handling missing values, encoding categorical variables, scaling numerical features and splitting the data into training and testing sets.

We removed all the missing values from the dataset by filling them or by removing the unnecessary columns.

We split the dataset into feature variables, consisting of meteorological parameters, and target variables, representing next-day minimum and maximum temperatures. Afterward, we perform feature scaling to normalize the data distribution and prevent dominance of certain features during model training.

1. **Building machine learning model:**

**Model Training:**

We train the models in two steps, first for predicting next-day minimum temperature (Next\_Tmin) and then for predicting next-day maximum temperature (Next\_Tmax). We utilize various regression algorithms, including RandomForestRegressor, DecisionTreeRegressor, ExtraTreesRegressor, BaggingRegressor, AdaBoostRegressor, GradientBoostingRegressor, and XGBRegressor. Each model is trained and evaluated using cross-validation and performance metrics such as mean squared error and R-squared.

**Hyperparameter Tuning:**

To optimize the model's performance, we perform hyperparameter tuning using GridSearchCV. This technique systematically searches for the best combination of hyperparameters within predefined ranges. By fine-tuning the models' parameters, we aim to enhance their predictive accuracy and generalization capabilities.

**Model Evaluation:**

We evaluate the trained models using cross-validation scores and R-squared values to assess their performance on unseen data. The models with the highest cross-validation scores and R-squared values are considered optimal for predicting next-day temperatures.

**Prediction and Model Deployment:**

Finally, we save the best-performing model for next-day temperature prediction based on the results of hyperparameter tuning. This model can be deployed in production environments to make real-time predictions of next-day temperatures, enabling informed decision-making in various domains.

1. **Concluding statement:**

In this project, we demonstrated the use of machine learning for temperature forecasting. We explored dataset through EDA,pre-processed the data, and built several models to predict temperature. Our result indicates that machine learning models can effectively forecast temperature, with random forest showing improved performance over simple linear regression.

This project aims to improve the accuracy of next day & present maximum and minimum air temperature forecasts by correcting biases in the predictions generated by the LDAPS model. By implementing bias correction techniques and evaluating model performance, we seek to enhance the reliability of temperature forecasts, which can have significant implications for various sectors including agriculture, energy, and public safety.