**TEMPERATURE FORECAST PROJECT**

****

**Problem Definition**

Accurate weather prediction is an important aspect in various sectors like agriculture, disaster management and everyday life. Temperature fluctuations affect the living and non-living components of our environment equally. Understanding and forecasting these fluctuations can lead to better decision-making in numerous applications. For example, in agriculture, precise temperature forecasts help farmers plan their harvesting schedules and protect crops from adverse weather conditions. In disaster management, accurate weather predictions are important for preparing for extreme weather events and minimizing the impact of natural disasters. Additionally, in everyday life, reliable and accurate temperature forecasts guide people in their daily activities, from dressing appropriately for the weather to planning outdoor events.

So, the objective of this project is to predict the minimum and maximum temperatures for the next day using the weather dataset. The data was generated 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 from 2013-2017. Separate models need to be built 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.

**Data Analysis**

The dataset contained the following features:

*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*

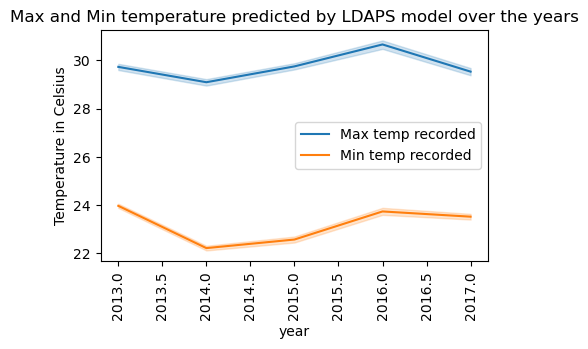
*24. Next\_Tmax - The next-day maximum air temperature (Â°C): 17.4 to 38.9*

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

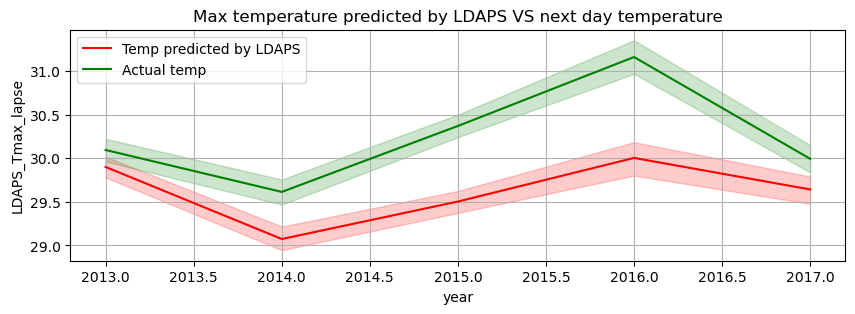
Data analysis was done using various python libraries such as pandas, matplotlib, and seaborn. Descriptive analysis of data was performed using pandas. Graphical analysis of data (lineplot, barplot, distplot, boxplot and heatmap) was done using matplotlib and seaborn.

The findings are listed below:-

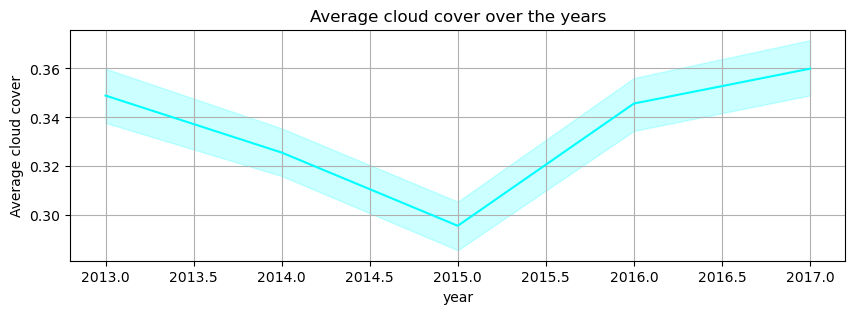
**OVERALL ANALYSIS:**



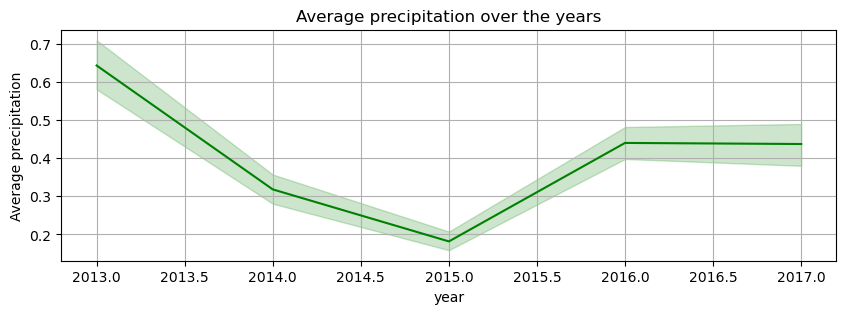
\* Over the years, maximum temperature was recorded on August,2016 and minimum on June, 2014



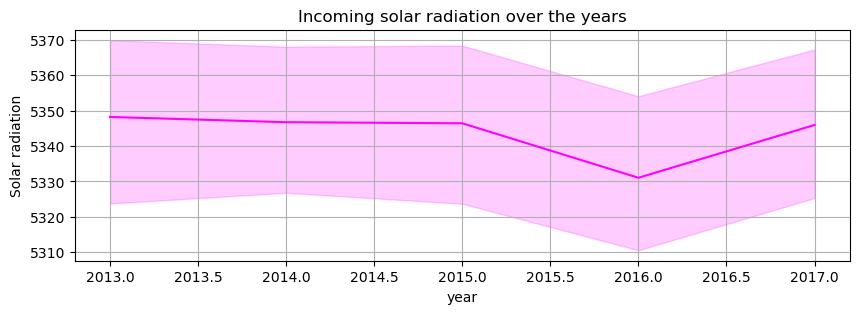
\* The temperature forecasted by the LDAPS model was close to the actual temperature recorded during 2013-2014 and later started varying. Much difference was recorded in 201-2016 and it almost started stabilizing by the beginning of 2017.



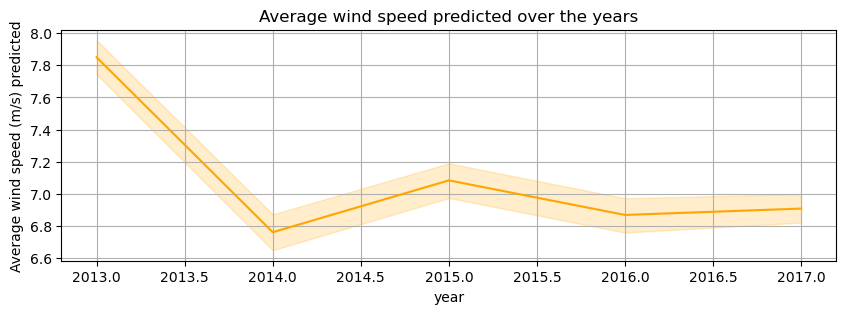
\* Average cloud cover forecasted by LDAPS model was highest in 2013 and lowest in 2015



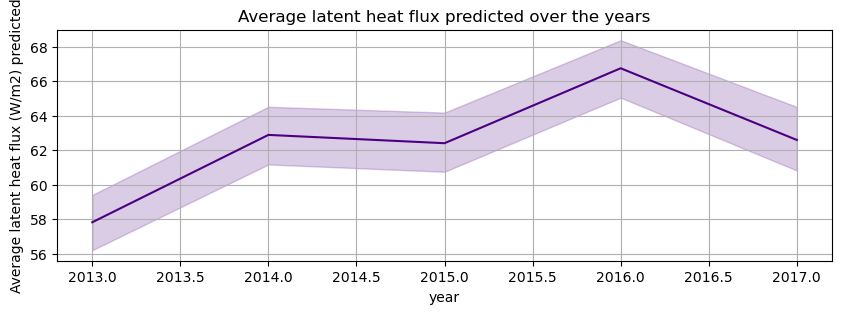
\* Average precipitation forecasted by LDAPS model was highest in 2013 and lowest in 2015



\* Solar radiation was almost equally received across the years except during 2015 it started declining and 2016 received the lowest radiation



\* Wind speed was highest in 2013 after which there was a steep decline in 2014



\* Latent heat flux was least in 2013 and peaked during 2016

\* Highest temp of '37.6°C' was recorded at station '18' on 11-08-2016.

\* Minimum temperature of the day was '26.8°C'

\* Location co-ordinates of the area/station which recorded the highest temperature is 37.4832,127.024 and the slope of the station is 1.2313 Â°

\* Average cloud cover was 31.75% and average precipitation was 0.0 indicating no rainfall on 11-08-2016

\* Correlation analysis revealed that previous day temperatures influence the temperature for the next day as they are positively correlated.

\* As cloud cover and precipitation are more, temperature tends to decrease and relative humidity tends to increase as per the correlation analysis.

**YEAR-WISE ANALYSIS**:

\* Over the years, maximum temperature was recorded on August,2016 and minimum on June, 2014

\* August 2016 had high average temperature range (24.57 to 32.14°C) and high incoming solar radiation of 4941.71 W/m², which can explain the hottest temperature during that period.

\* June 2014 had low average temperature range (20.87 to 29.48°C) and high incoming solar radiation of 5874.69 W/m². The low temperature makes it the coolest period in the whole dataset. Low cloud cover (7.59%) and low precipitation (0.72%) rate can explain the high incoming solar radiation during that period.

\* The LDAPS model was able to predict the next day temperatures successfully in the beginning but significant deviations were observed from late 2014 to 2016 which can be due to the changing weather patterns or due to the limitations of the model.

**EDA Concluding Remarks**

Exploratory Data Analysis (EDA) was done to understand the dataset and uncover hidden patterns in the data. Here are some key remarks derived from the EDA:

1. **Temperature Trends:** Over the years, the highest maximum temperature was recorded in August 2016, while the lowest minimum temperature was recorded in June 2014. This indicates significant seasonal variations in temperature.

2. **LDAPS Model Accuracy:** The temperature forecasted by the LDAPS model closely matched the actual temperature in the early years (2013-2014). However, deviations increased significantly from 2015 to 2016, suggesting limitations in the predictive capabilities of the model during these years.

3. **Cloud Cover and Precipitation**: The average cloud cover and precipitation forecasted by the LDAPS model were highest in 2013 and lowest in 2015.

4. **Solar Radiation and Wind Speed**: Solar radiation remained consistent across the years, except for a decline in 2015 and the lowest radiation was received in 2016. Wind speed was highest in 2013, with a steep decline in 2014, indicating changes in wind patterns over time.

5. **Latent Heat Flux**: Latent heat flux was lowest in 2013 and peaked in 2016.

6. **Correlation Analysis:** Correlation analysis revealed that temperatures from the previous day are positively correlated with the next day's temperatures. Higher cloud cover and precipitation are associated with lower temperatures, while higher relative humidity tends to increase temperature.

**Pre-processing Pipeline**

Data pre-processing is a critical step to ensure that good quality data is fed to the machine learning algorithms to get accurate predictions.

The following libraries were imported for the project:

# general libraries

* *import numpy as np*
* *import pandas as pd*
* *import matplotlib.pyplot as plt*
* *import seaborn as sns*
* *import pickle*
* *import warnings*
* *warnings.filterwarnings('ignore')*

# data pre-processing libraries

* *from sklearn.model\_selection import train\_test\_split*
* *from sklearn.preprocessing import LabelEncoder*
* *from sklearn.preprocessing import StandardScaler*

# feature selection libraries

* *from statsmodels.stats.outliers\_influence import variance\_inflation\_factor*
* *from sklearn.feature\_selection import SelectKBest, f\_classif*

The following pre-processing steps were done with the dataset:

**Data cleaning –**

* Removal of unwanted columns was done using pandas library
* Nulls were dropped using dropna() method as several records had nulls across multiple columns
* Outlier removal was done using scipy.stats.zscore
* There is significant left skewness in ppt column and small left skewness in DEM, Slope and LDAPS\_WS as evident from the histogram. Skewness was quantified using skew() and managed using cuberoot function of numpy library
* Number of rows removed after cleaning and pre-processing is 164

**Feature Engineering –**

* Categorical columns were encoded using LabelEncoder() of sklearn.preprocessing library
* Creation of new columns such as year, month and location co-ordinates for data analysis and modelling purpose using pandas
* Two new columns were created after taking average of all 4 quarter wise data in the case of precipitation and cloud cover
* VIF scores were calculated for all the feature columns using *statsmodels.stats.outliers\_influence.variance\_inflation\_factor* to find if any columns are multicollinear. Multicollinearity occurs when two or more independent features are highly correlated with each other which makes the model less efficient.
* SelectKBest was used to screen out the top 10 features for training the model.
* Scaling of data to ensure that all numerical features were on a similar scale using sklearn.preprocessing.StandardScaler
* After scaling, the data is split into train and test sets using train\_test\_split function prior to model training

**Building Machine Learning Models**

The following libraries were imported for the model training and evaluation:

# import algorithms

* *from sklearn.linear\_model import LinearRegression*
* *from sklearn.linear\_model import Lasso*
* *from sklearn.linear\_model import Ridge*
* *from sklearn.tree import DecisionTreeRegressor*
* *from sklearn.ensemble import RandomForestRegressor*
* *import xgboost as xgb*
* *from sklearn.svm import SVR*

# model evaluation metrics

* *from sklearn import metrics*
* *from sklearn.metrics import accuracy\_score, confusion\_matrix, roc\_curve, roc\_auc\_score,classification\_report*
* *from sklearn.model\_selection import cross\_val\_score*

Seven models were trained for the regression problems including LinearRegression, Lasso, Ridge, DecisionTreeRegressor, RandomForestRegressor, XGBRegressor, and Support Vector Machine. These models were evaluated using appropriate metrics from sklearn.metrics to find the best model: r2 score, Mean Squared Error, Mean Absolute Error and Cross Validation. By comparing these metrics, the analysis determined the most effective models for predicting temperature variations and validated the robustness of these models against different evaluation criteria. This approach ensured that the selected models not only performed well on the training data but also generalized effectively to unseen data, providing a reliable foundation for future temperature forecasting.

**Predicting Minimum Temperature**

Out of the 7 models trained, the XGBoost model was found to be the best performer for predicting the minimum temperature for the next day with 91% accuracy. Here are the evaluation metrics for the XGBoost model:

Test R2 score: 0.91

Train R2 score: 0.99

Cross Validation: 0.77

MSE: 0.46

MAE: 0.52

The high R2 scores and low MSE and MAE values indicate that the XGBoost model accurately predicts the minimum temperature with minimal overfitting. The cross-validation score further supports the model's efficiency.

**Predicting Maximum Temperature**

Out of the 7 models trained, the Random Forest model was found to be the best performer for predicting the maximum temperature for the next day with 98% accuracy. Here are the evaluation metrics for the Random Forest model:

Test R2 score: 0.98

Train R2 score: 0.88

Cross Validation: 0.68

MSE: 1.13

MAE: 0.79

The Random Forest model achieved a high test R2 score, indicating excellent predictive performance. The train R2 score, though slightly lower, suggests minimal overfitting. The cross-validation score, while not as high as the test R2 score, still demonstrates the model's generalizability.

**Concluding Remarks**

Exploratory Data Analysis (EDA) revealed significant seasonal variations in the data generated during 2013-2017, with the highest maximum temperature recorded in August 2016 and the lowest minimum temperature in June 2014. The accuracy of the LDAPS model was good in the early years, closely matching actual temperatures from 2013 to 2014, but its predictions showed variations from 2015 to 2016. Cloud cover and precipitation were highest in 2013 and lowest in 2015, which affected solar radiation patterns; solar radiation remained steady overall but declined in 2015 and was at its lowest in 2016. Wind speed peaked in 2013 and declined sharply in 2014, highlighting changing wind patterns, while latent heat flux was lowest in 2013 and peaked in 2016. Correlation analysis confirmed that previous day temperatures strongly influence the next day's temperatures, with higher cloud cover and precipitation linked to lower temperatures and higher humidity associated with increased temperatures. The EDA and model evaluation underscore the effectiveness of machine learning algorithms, particularly XGBoost and Random Forest, in accurately predicting daily temperature variations. These models can improve weather forecasting and provide valuable insights for various applications, such as agricultural planning and disaster management.