Dataset Description

1 Dataset

The description of the variables involved in the dataset are given below. This may help you interpret the dependency between variables and the model results.

- **Mean sea level pressure (msl):** Pressure of the atmosphere at the surface of the Earth, adjusted to the height of mean sea level. (Unit - Pa)
- **Sea surface temperature (sst):** Temperature of sea water near the surface. (Unit - K)
- **10m u-component of wind (u10):** Horizontal speed of air moving towards the east, at a height of 10m above the surface of the Earth. (Unit - m/s)
- **10m v-component of wind (v10):** Horizontal speed of air moving towards the north, at a height of 10m above the surface of the Earth. (Unit - m/s)
- **Wind speed (ws):** Overall speed at which the air is moving. (Unit - m/s)
- **2m temperature (t2m):** Temperature of air at 2m above the surface of land, sea or inland waters. (Unit - K)
- **Relative Humidity (rh):** Amount of water vapor present in the air relative to the maximum amount it can hold at a specific temperature and pressure (Unit - %)
- **Total precipitation (tp):** Accumulated liquid and frozen water, comprising rain and snow, that falls to the Earth's surface. (Unit - m)

2 Exploratory Data Analysis

- Descriptive Statistics:
  - Get the summary statistics of numerical variables.
  - Plot histograms for numerical variables, observe the distribution and interpret the distribution shapes (e.g., normal, skewed, bimodal).
  - Calculate the mean and variance for numerical variables ignoring the NA values if any.
- Find Relationships Between Variables:

  - Visualize relationships between numeric variables (use scatter plots, histograms, box plots, etc).
  - Compute the correlation matrix and plot.
  - Discuss any strong correlations, noticeable or interesting patterns that you observe.

3 Model

Build the linear regression (LR) model for the target variable “tp ”

- Split your data into train and test sets (80-20%).
- Fit the model on the training data and report both train and test Mean Square Error.
- Interpret the model coefficients.
- Does the performance of LR model improve with the removal of some of the highly correlated variables?
- Use non-linear features
- Try Ridge regression and LASSO regression.