Weather Forecasting

You are a data scientist working for a smart agriculture startup. Farmers rely on accurate weather predictions to plan irrigation, planting, and harvesting. However, traditional weather forecasts are not always reliable for hyper-local conditions. Your task is to build a machine learning model that predicts whether it will rain or not based on historical weather data. This data has been manually entered into an Excel sheet, but there are some mistakes, such as missing values, incorrect entries, and formatting inconsistencies, which need to be cleaned and processed before training the model.

The dataset provided contains daily weather observations for 300 days, including:

- **avg_temperature**: Average temperature in °C
- **humidity**: Humidity in percentage
- **avg_wind_speed**: Average wind speed in km/h
- **rain_or_not**: Binary label (1 = rain, 0 = no rain)
- **date**: Date of observation

Your goal is to predict the `rain_or_not` label for future 21 days based on the given features.

Part - 1 - 85%

Participants are expected to:

1. Preprocess the dataset (e.g., handle missing values, encode features, etc.). - 20%
2. Perform exploratory data analysis (EDA) to understand relationships between features and the target variable. - 30%
3. Train and evaluate machine learning models (e.g., logistic regression, decision trees, random forests, gradient boosting, etc.). - 20 %
4. Optimize the model using hyperparameter tuning and feature engineering. -10%
5. final output should able to provide the probability of rain - 5%

Part - 2 - 15%

You are a MLOps Engineer of this startup assuming API developers are able to get real time data within 1 minute intervals.your goal to design a system to present for the project manager to predict the next 21 daily basis raining probability. These IOT devices and sensors are malfunctioning sometimes and the system should be able to handle these.

Participants are expected to:

1. Design System diagram (Not mandatory to use specific symbols) that shows clearly the data flows from source to user

2. Simple description of each component

Participants Should Deliver

1. Source JupyterNotebook Script
2. Report-1 (do not exceed 15 pages except cover) for part-1
3. Report-2 (do not exceed 3 pages except cover) for part-2

Download the dataset from here