**Hackathon Report**

**Team Name:** Team Hawks

**Problem Definition:** Use weather datasets to predict temperature, rainfall, or other conditions for specific regions. This can help in planning for agricultural or travel needs

**Aim:** Weather Data Analysis and Prediction

### **Datasets Overview**

#### **1. Dataset 1:** [weather\_data](https://docs.google.com/spreadsheets/d/1U3_PY6l4tmc285OujdxrU5L_PzAmpsMZ9OBJhT4DRbw/edit?gid=206377849#gid=206377849)

#### **2. Dataset 2:** [weather\_data\_states](https://docs.google.com/spreadsheets/d/1KErABQXsGaR-xC3ofyUpphThwr6RSUBLO32rDLl2XOo/edit?gid=1362299547#gid=1362299547)

#### **3. Dataset 3:** [Weather\_dataset](https://docs.google.com/spreadsheets/d/1XToyR5NzmTZw1Fw7tusagoJQyMIRGljk07iATjR1VmE/edit?gid=0#gid=0)

### **Data Cleaning and Solution Planning**

**Data Cleaning:** Three datasets were collected through data scraping and insights derived from websites such as timeanddate.com and worldweatheronline.com. These datasets provided diverse information, including time, date, and weather trends.  
  
The collected data was processed using Python's **pandas** library and Excel tools. The steps included:

* **Data Viewing:** Initial examination of datasets to understand structure and content.
* **Handling Missing Observations (NA):** Addressing missing values using appropriate imputation techniques or removal.
* **Typecasting:** Ensuring data types align with expected formats for analysis.
* **Summary Statistics:** Generating descriptive statistics to understand key metrics.

**Solution Planning:**Following are the steps involved in the solution process:  
 **Exploratory Data Analysis (EDA):**

* A systematic examination of the datasets to identify trends, patterns, and anomalies. As per weather data, historical trends are examined and descriptions of climatic conditions we derived for comparative analysis.

**Descriptive Analysis:**

* Calculation of key measures such as mean, median, and mode to summarize central tendencies.
* Average of Temperatures, Percentage division for description, Date wise calculation of trends.

**Distribution Analysis:**

* Understanding the spread and behavior of data across different fields.  
  Wind speed, Rainy trends, sunny weather etc. Fields like these are distributed as per required part of data per calculations.

**Data Visualization:**

* Creating charts and plots (e.g., histograms, scatter plots, box plots) to visually represent insights from the data.

**Tools Used:**R programming  
PowerBi  
Python: pandas, Matplot, Seaborn

### **Methodology**

The following machine learning techniques were applied :

1. **Logistic Regression:**
   * A statistical model used for binary classification problems, evaluating relationships between input features and output labels. It predicts the probability of an event, such as whether it will rain or not, based on input features like temperature, humidity, wind speed, and atmospheric pressure. These probabilities can then be converted into categorical predictions (e.g., "Rain" or "No Rain"). In weather prediction, it is especially valuable when combined with proper feature selection and preprocessing to handle complex meteorological data.
2. **Decision Trees:**
   * A tree-structured model for decision-making and classification, offering interpretable outputs.
3. **Random Forest:**
   * An ensemble method combining multiple decision trees to improve prediction accuracy and robustness.  
     This algorithm will ensemble learning method that builds multiple decision trees and combines their outputs to make more accurate and robust predictions for weather .
4. **Hyperparameter Tuning:**
   * Optimization of model parameters to enhance performance using techniques such as grid search or random search.

### **Model Development Steps**

1. **Model Evaluation:**
   * Assessing the performance of models using predefined metrics.
2. **Confusion Matrix:**
   * A tabular representation of actual vs. predicted outcomes to evaluate classification performance.  
     Example: A performance evaluation tool used in classification models, including weather prediction. It helps analyze how well the model predicts different classes, such as "Rain" and "No Rain.
3. **Accuracy:**
   * Proportion of correctly predicted instances to total instances. Historical data is compared with the predicted ones.
4. **Precision:**
   * The ratio of true positives to the sum of true positives and false positives.
5. **R-Squared Value:**
   * A measure of how well the model explains the variance in the data.  
     In weather data analysis, **R-squared** is used to statisticaly measure used to evaluate the goodness-of-fit of a regression model, indicating how well the model's predictions align with the actual observed data.
6. **F-Value Calculation:**
   * Combining precision and recall into a single performance metric, providing a balanced evaluation of the model's effectiveness.