Project Documentation: Algerian Forest Fires FWI Prediction

1. Introduction

This project focuses on predicting the **Fire Weather Index (FWI)**, a key measure of fire danger, using meteorological and FWI component data from the Algerian Forest Fires dataset. The FWI system, developed by the Canadian Forest Service, provides a daily numerical rating of fire danger. Predicting this index helps fire management agencies in proactive planning, resource allocation, and wildfire prevention efforts.

2. Dataset Description

The dataset contains **244 instances** of daily forest fire danger ratings from **June to September 2012**, covering two distinct regions in Algeria: **Bejaia** (northeast) and **Sidi Bel-abbes** (northwest).

Column Name	Description	Units/Scale
Day, Month, Year	Date of observation (temporal features).	Day, Month (1-12), Year (2012)
Temperature (Temp)	Maximum noon temperature.	°C
Relative Humidity (RH)	Relative Humidity.	%
Wind Speed (Ws)	Wind speed.	km/h
Rain	Total daily rainfall.	mm
Fine Fuel Moisture Code (FFMC)	Index related to the moisture content of fine fuels (litter, small branches).	0 to 100
Duff Moisture Code (DMC)	Index related to the moisture content of loosely compacted organic material (duff) beneath the litter.	Numeric Index
Drought Code (DC)	Index related to the moisture content of deep organic layers.	Numeric Index
Initial Spread Index (ISI)	Index related to the expected rate of fire spread.	Numeric Index

Column Name	Description	Units/Scale
Buildup Index (BUI)	Index related to the total amount of fuel available.	Numeric Index
FWI	**Fire Weather Index (Target Variable)**. A numerical index representing the intensity of a fire.	Numeric Index
Classes	Categorical class indicating if a fire occurred ('fire', 'not fire').	Categorical
Region	Region identifier (encoded as 0 or 1).	Categorical

3. Methodology

3.1 Data Loading and Initial Inspection

The project begins by loading the Algerian_forest_fires_dataset.csv file using **pandas**. Initial checks are performed to understand the dataset structure:

- **Head:** Inspecting the first 5 rows.
- **Shape:** Determining the number of rows and columns.
- **Columns:** Listing all column names.
- **Info:** Checking data types and memory usage.
- **Missing Values:** Summing NaN values per column.

3.2 Data Preprocessing and Cleaning

1. **Duplicate Removal:** Duplicate rows are dropped to ensure the analysis is based on unique daily observations.
    - *Code:* df.drop_duplicates(inplace=True)
2. **Categorical Encoding:** The categorical columns, including the region and fire classes (likely 'Classes' and 'Region' in the original data), are converted into numerical representations using **label encoding** (.astype('category').cat.codes).
3. **Summary Statistics:** Descriptive statistics are calculated for all remaining numerical columns to understand data distribution, central tendency, and spread.

4. **Cleaned Data Export:** The processed DataFrame is saved as cleaned_fwi_dataset.csv for use in subsequent model training phases.

3.3 Exploratory Data Analysis (EDA)

3.3.1 Univariate Analysis (Histogram)

A histogram is generated for the selected column (e.g., **Temperature**, **RH**, **FWI**) to visualize its distribution. This helps in understanding the frequency of different values and identifying potential skewness or outliers.

- *Tool:* **Matplotlib** (plt.hist)

3.3.2 Multivariate Analysis (Correlation Heatmap)

A **correlation heatmap** is created to visualize the linear relationship between all numerical features. The goal is to identify which features are most strongly correlated with the **Target Variable (FWI)**, as these will be the most influential predictors in the regression model.

- *Tool:* **Seaborn** (sns.heatmap)
- *Interpretation:* High correlation (close to 1 or -1) indicates a strong relationship, suggesting that a change in one variable is consistently associated with a change in the other.