Problem statement and datasets

Form a group of 2 people from different nationalities. Identify **3 possible datasets** that you plan to use to solve the problem for your class project. From the problem statement and datasets, answer the following questions.

Problem: Briefly explain the problem that you plan to solve using the data science process, possible users, and its impact.

Dataset description: Explain details of the datasets you select e.g. what's it about? How many features, and how many records?

Justification: Justify that your data has enough information to solve your problem.

Don't forget to properly cite all datasets, and pictures that you use to answer all questions. Datasets can complement each other to solve your problem.

Some places you can find datasets on:
https://archive.ics.uci.edu/
https://data.gov/
https://datacatalog.worldbank.org/home
https://old.datahub.io/dataset
https://github.com/awesomedata/awesome-public-datasets

Submitted by:
Student Name: Sachin Malego
Student ID: 125171
Student Name: Sila N Mahoot
Student ID: 125127

Problem Statement: Briefly explain the problem that you plan to solve using the data science process, possible users, and its impact.

A preliminary study to understand the impact of El Nino on Forest Fire leading to poor Air Quality (PM2.5) using Python and Machine Learning.

Forest fires are one of the prominent environmental and socio-economic challenge with wide-ranging impacts on biodiversity, economics, and public health. The increasing frequency of forest fires around the globe have been exacerbated by changing climatic conditions, leading to significant environmental degradation, economic losses, and severe health issues due to deteriorating air quality. Unusual El Nino events disrupts normal weather patterns globally, creating conditions that can exacerbate the risk of forest fires in vulnerable regions. Thus, the cascading effects of climate change are closely linked with the El Nino-Southern Oscillation (ENSO) phenomenon being a critical driver of these events.

Despite the known association between El Nino and extreme weather events, the specific impact of El Nino on forest fire and its resulting impact on air quality has not been thoroughly quantified

or integrated into predictive models. This prominent gap in research limits the ability of the environment agencies, policymakers, and emergency services to anticipate and mitigate the effects of forest fires effectively.

Now, the **problem that we plan to solve** involves developing a preliminary model to analyze past forest fire occurrences based on climatic conditions, with a specific focus on understanding the impact of El Nino on these events. By developing a model that uses climate data, including El Nino indicators, we aim to forecast the likelihood and cascading effects of these events in cause of forest fire and air quality. Our objective from this is to create a data-driven model that leverages historical climate data, including El Nino indicators, to predict the impacts and likelihood of forest fires in vulnerable regions and its possible effects in air quality and public health.

This model can serve as a tool for various **stakeholders**, including but not limited to Environmental Agencies, Policymakers, Emergency Services, and Researchers and Scientists. This will support these stakeholders to monitor and assess fire risks, make informed decision regarding resource allocation, policy formulation, and the development of fire management strategies. Furthermore, it can help emergency services to strategically plan and deploy resources, ensuring rapid response to potential fire outbreaks. Additionally, researchers and scientists can use this preliminary model and further enhance it contributing to the broader understanding of environmental dynamics.

The **impact** of this project although preliminary can be far outreaching. This preliminary model can help infer causes of extreme events and help mitigate the adverse effects on ecosystem, reduce economic losses associated with fire damage, and protect public health by preventing the deterioration of air quality. Additionally, the insights gained from this project will enhance our understanding of how global climatic phenomena like El Nino contribute to environmental hazards, providing a foundation for future research and policy development.

Dataset description: Explain details of the datasets you select e.g. what's it about? How many features, and how many records?

To make an **inference on the problem statement** stated above we have decided to use **3 datasets** from three different sources of climate agencies. Using these datasets, we aim to analyze forest fires occurrences based on climatic conditions, particularly the influence of El Nino events. These datasets are crucial for understanding and modelling the factors contributing to forest fires.

Dataset 1: **El Nino (*https://archive.ics.uci.edu/dataset/122/el+nino*)**

El Nino dataset is related to the oceanographic and surface meteorological readings collected from a series of buoys positioned throughout the equatorial Pacific. This dataset helps monitor and understand climate variations, particularly those related to the El Nino/Southern Oscillation (ENSO) cycles. The data helps in predicting weather conditions and studying the relationships between various climate-related variables. It's critical for understanding seasonal-to-interannual climate variations, especially those originating in the tropics. The data comes from nearly 70 moored buoys that measure air temperature, relative humidity, surface winds, sea surface temperatures, and subsurface temperatures down to 500 meters. Some buoys also measure currents, rainfall, and solar radiation.

Dataset Characteristics:

- **Number of Records (Instances)**: 178,080
- **Number of Features (Variables)**: 11, which include both continuous and integer variables.

Features:

- **ID**: Unique identifier for each record.
- **Year**: The year when the data was recorded.
- **Month**: The month when the data was recorded.
- **Day**: The day when the data was recorded.
- **Date**: Full date of the record.
- **Latitude**: The latitude of the buoy.
- **Longitude**: The longitude of the buoy.
- **Zonal Winds**: Wind speed in the east-west direction (west<0, east>0).
- **Meridional Winds**: Wind speed in the north-south direction (south<0, north>0).
- **Humidity**: Relative humidity percentage.
- **Air Temperature**: Temperature of the air at the location.
- **Sea Surface Temperature (SS Temp)**: Temperature of the sea surface.

Missing Values:

- The dataset does contain missing values, represented by periods (.), which are due to some buoys not being equipped to measure certain variables like currents, rainfall, or solar radiation.

Dataset 2: **Forest Fires (*https://archive.ics.uci.edu/dataset/162/forest+fires*)**

Forest fires dataset is related to the task of predicting the burned area of forest fires in the northeast region of Portugal. This dataset is particularly useful for studying the impact of meteorological and environmental conditions on forest fire occurrences and severity.

Dataset Characteristics

- **Number of Instances:** 517
- **Number of Features:** 12
- **Associated Task:** Regression
- **Feature Type:** Real

Features

- **X (Feature)**
  Type: Integer
  Description: X-axis spatial coordinate (values ranging from 1 to 9).
- **Y (Feature)**
  Type: Integer
  Description: Y-axis spatial coordinate (values ranging from 2 to 9).
- **month (Feature)**
  Type: Categorical
  Description: Month of the year, with possible values ranging from 'Jan' to 'Dec'.
- **day (Feature)**

Type: Categorical
Description: Day of the week, with possible values ranging from 'Mon' to 'Sun'.

- **FFMC (Feature)**
  Type: Continuous
  Description: Fine Fuel Moisture Code index from the Fire Weather Index (values ranging from 18.7 to 96.20).
- **DMC (Feature)**
  Type: Integer
  Description: Duff Moisture Code index from the Fire Weather Index (values ranging from 1.1 to 291.3).
- **DC (Feature)**
  Type: Continuous
  Description: Drought Code index from the Fire Weather Index (values ranging from 7.9 to 860.6).
- **ISI (Feature)**
  Type: Continuous
  Description: Initial Spread Index from the Fire Weather Index (values ranging from 0.0 to 56.10).
- **temp (Feature)**
  Type: Continuous
  Description: Temperature in degrees Celsius (values ranging from 2.2 to 33.30).
- **RH (Feature)**
  Type: Integer
  Description: Relative humidity as a percentage (values ranging from 15.0 to 100).
- **wind (Feature)**
  Type: Continuous
  Description: Wind speed in kilometers per hour (values ranging from 0.40 to 9.40).
- **rain (Feature)**
  Type: Integer
  Description: Rainfall in millimeters per square meter (values ranging from 0.0 to 6.4).

Target Variable

- **area (Target)**
  Type: Integer
  Description: Burned area of the forest in hectares (values ranging from 0.00 to 1090.84).

This target variable is highly skewed towards 0.0. It is advisable to apply a logarithmic transformation (ln(x+1)) to address this skewness and improve model performance.

Dataset 3: **Beijing PM2.5 (*https://archive.ics.uci.edu/dataset/381/beijing+pm2+5+data*)**
Beijing PM2.5 dataset is related to assess the influence of meteorological conditions on PM2.5 pollution levels in Beijing. Through statistical analysis, the dataset enables the quantification of PM2.5 severity and evaluates the efficacy of pollution reduction targets set by China's State Council. Analysis of the adjusted PM2.5 levels and associated percentiles reveals significant increases in PM2.5 concentrations during the years 2013 and 2014 compared to 2012. This dataset provides a comprehensive foundation for exploring the dynamics of air pollution and the

effectiveness of environmental policies. The integration of PM2.5 data with meteorological variables facilitates a nuanced understanding of the factors contributing to air quality variations and informs policy adjustments for improved air quality management.

This dataset integrates hourly PM2.5 measurements with various meteorological parameters such as temperature, dew point, atmospheric pressure, wind speed, and precipitation metrics.

Dataset Characteristics
Type: Multivariate Time-Series
Temporal Scope: January 1, 2010, to December 31, 2014
Number of Instances: 43,824 hourly observations
Number of Features: 11

Features

- **No**: Row number (integer, non-missing)
- **year**: Year of observation (integer, non-missing)
- **month**: Month of observation (integer, non-missing)
- **day**: Day of observation (integer, non-missing)
- **hour**: Hour of observation (integer, non-missing)
- **pm2.5**: PM2.5 concentration (μg/m³), target variable (integer, missing values present)
- **DEWP**: Dew Point (°C) (integer, non-missing)
- **TEMP**: Temperature (°C) (integer, non-missing)
- **PRES**: Pressure (hPa) (integer, non-missing)
- **cbwd**: Combined wind direction (categorical, non-missing)
- **Iws**: Cumulated wind speed (m/s) (continuous, non-missing)
- **Is**: Cumulated hours of snow (integer, non-missing)
- **Ir**: Cumulated hours of rain (integer, non-missing)

Missing Values: The dataset contains missing values primarily in the pm2.5 feature, which denotes instances where PM2.5 concentration data is unavailable.

Associated Tasks: The primary task associated with this dataset is regression analysis, aimed at predicting PM2.5 concentrations based on meteorological conditions.

Justification: Justify that your data has enough information to solve your problem.
To understand the intricate relationship between the climatic conditions, forest fires, and air quality we have selected three comprehensive datasets that collectively provides a foundation to address the problem of analyzing the impact of El Nino on occurrence of climatic events like the forest fire and the resulting air quality (PM2.5).

1. **El Nino dataset**
   This dataset provides information on El Nino events, which are central to understanding the impact of these climatic anomalies on forest fire. This dataset contains 178,080 records and 11 features of temporal and spatial prespective providing an analysis of how El Nino conditions vary and how these variations correlate with forest fire occurrences. The

presence of missing values, while present, is manageable. Imputation techniques can be applied to handle these missing values.

2. **Forest Fires dataset**
   This dataset specifically focuses on forest fires in the northeast region of Portugal. The features such as Fine Fuel Moisture Code, Duff Moisture Code, and Drought Code provide direct indicators of fire risk and intensity. With 517 instances and 12 features, this dataset includes critical variables that describe the conditions leading to forest fires. Although the dataset is limited in size compared to others, it provides detailed information about fire incidents, which is essential for understanding how different factors contribute to fire severity.

3. **Beijing PM2.5 dataset**
   This dataset provides insights into air quality, specifically PM2.5 concentrations, which are directly affected by forest fires. With 43,824 hourly observations and 11 features, this dataset offers a detailed view of air quality over a substantial period (2010-2014). The inclusion of meteorological parameters such as temperature, dew point, and wind speed support the analysis by providing context for variations in PM2.5 levels. Although there are few missing values these can be addressed using imputation techniques to ensure a complete dataset for analysis.

Conclusion

The combined use of these datasets provides an articulate approach to analyzing the impact of El Nino on forest fires and air quality. The El Nino dataset offers insights into climatic conditions related to the phenomenon, the forest fires dataset provides detailed information on fire occurrences and intensity, and the PM2.5 dataset enables the evaluation of air quality impacts. Together, these datasets enable the development of a preliminary model, addressing the problem from multiple perspectives and offering valuable insights for environmental agencies, policymakers, and emergency services.

Citation

[1] UCI Machine Learning Repository. Available at: https://archive.ics.uci.edu/dataset/122/el+nino (Accessed: 23 August 2024).

[2] Cortez, P. and Morais, A. (2007) *Forest Fires Dataset*. Available at: https://archive.ics.uci.edu/dataset/162/forest+fires (Accessed: 23 August 2024).

[3] UCI Machine Learning Repository. Available at: https://archive.ics.uci.edu/dataset/381/beijing+pm2+5+data (Accessed: 23 August 2024).