DATA MINING

Practical Work Project

Forest fires represent a major environmental and socio-economic challenge, leading to the loss of vegetation, soil degradation, and severe ecological damage. Early prediction of fire occurrence is therefore essential for effective fire prevention and management strategies.

In this project, we aim to develop a **data-driven predictive model** that uses **soil characteristics** (such as moisture, organic content, and texture) and **climate data** (temperature, humidity, rainfall, wind speed, etc.) to forecast the likelihood of forest fires. By applying **data mining and machine learning techniques**, we will explore patterns and relationships between environmental variables and fire events. The insights generated could support the design of **early warning systems** and **fire prevention plans**.

The main objectives of this project are to:

- Collect and preprocess soil and climate data relevant to fire prediction.
- Apply **supervised learning algorithms** to predict fire occurrence.
- Use **unsupervised learning (clustering)** to identify natural groupings and high-risk areas.
- Evaluate model performance using standard metrics.
- Provide interpretable insights on **fire risk zones**.

The project follows a three-phase data mining approach combining analysis, prediction, and clustering.

Step 1: Data Analysis and Preprocessing

- Exploratory Data Analysis
- Data Preprocessing
- Data Integration.
- Feature Engineering

Step 2: Supervised Machine Learning Algorithms

- From scratch development of K-Nearest Neighbors (KNN)
- From scratch development of Decision Trees
- From scratch development of Random Forest
- Evaluate the obtained models
- Comparative analysis with Scikit-learn implementations to assess performance.

Step 3: Unsupervised Machine Learning (Clustering)

- From scratch development of K-Means
- From scratch development of DBSCAN
- From scratch development of CLARANS
- Evaluate the obtained models
- Comparative analysis with Scikit-learn implementations to assess performance.

Study Area:

Algeria & Tunisia (grouped in the same dataset)
Climate and Fire: 2024

DATASETS:

Fire dataset:
Algeria & Tunisia: https://firms.modaps.eosdis.nasa.gov/country/
Instrument > VIIRS NOAA-20 > 2024 > Algeria
Instrument > VIIRS NOAA-20 > 2024 > Tunisia

Land Cover dataset:
Algeria: https://data.apps.fao.org/catalog/iso/0e958049-2a0a-4935-83c8-af78626068fc
Tunisia: https://data.apps.fao.org/catalog/iso/d0ba96c7-786c-4f3f-bbd9-e427b3b23d2d

Climate dataset:
WorldWide: https://worldclim.org/data/monthlywth.html at 5 minutes

Elevation dataset:
WorldWide:
https://edcintl.cr.usgs.gov/downloads/sciweb1/shared/topo/downloads/GMTED/Grid_ZipFiles/be15_grd.zip

Soil dataset:
WorldWide:
https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/harmonized-world-soil-database-v20/en/

- Get the features from Table: HWSD2_LAYERS
- Only keep the values of LAYER = D1 (which means the characteristics of the first 20cm depth of the soil.
- Extract the features: “COARSE”, “SAND”, “SILT”, “CLAY”, “TEXTURE_USDA”, “TEXTURE_SOTER”, BULK”, “REF_BULK”, “ORG_CARBON”, “PH_WATER”, “TOTAL_N”, “CN_RATIO”, “CEC_SOIL”, “CEC_CLAY”, “CEC_EFF”, “TEB”, “BSAT”, “ALUM_SAT”, “ESP”, “TCARBON_EQ”, “GYPSUM”, “ELEC_COND”.