Wildfire Prediction, Cause Classification, and Size Class Analysis

This repository implements machine learning models, including Random Forest and Light Gradient Boosting Machine, for predicting wildfire occurrence, classifying fire causes, and predicting the fire size class. The goal is to enhance early wildfire detection, identify contributing factors, and estimate fire severity for effective mitigation strategies.

Introduction

Traditional wildfire management systems often rely on static thresholds for environmental conditions, missing complex interactions between variables. This project leverages machine learning models on environmental and geospatial data to:

1. Predict wildfire occurrence.
2. Classify the causes of wildfires.
3. Predict the size class of fires to estimate potential severity and resource needs.

Dataset

The dataset integrates multiple sources to provide comprehensive insights for wildfire prediction and analysis:

1. **‘weather_readings’:**
    - Temperature, humidity, wind speed, precipitation, and other environmental conditions.
    - Data sourced from weather stations strategically placed across fire-prone areas.
2. **Trail Data:**
    - Locations of recreational trails frequently visited by people.
    - Used to analyze the impact of human activity on fire occurrences and causes (e.g., "recreation"). included in the EDA notebook directly.
3. **‘Oil and Gas Industry’:**
    - Locations of oil and gas operations, industrial facilities, and related infrastructure.
    - Explored as a contributing factor to fires categorized under "industry."

4. **‘processed_wildfire_df’:**
    - Historical data on fire occurrences, including:
        - General Causes: Lightning, industry, recreation, and residential.
        - Fire Size Classes: Small, medium, large, and others to indicate fire severity.
5. **‘Alberta outline.png:**
    - An equilinear map of Alberta.
6. **‘processed_weather_stations’:**
    - Locations and names of the stations with their latitude and longitude data and name of the forest areas.

The main datasets for weather readings, wildfire, and weather stations were provided by the professor and available on eclass.

Pipeline Overview

- **Exploratory Data Analysis (EDA):**
    - Analyze temporal, spatial, and weather-related patterns in the dataset.
    - Identify correlations between variables and fire occurrence, causes, and size classes.
- **Data Preprocessing:**
    - Handle missing values with median imputation.
    - Standardize numerical features.
    - Generate lagged features and rolling averages to capture temporal trends.
- **Modeling:**
    - **Task 1: Predict wildfire occurrence (binary classification).**
        - Train a Random Forest model using weather, terrain, and temporal features.
        - Evaluate using accuracy, precision, recall, and F1-score.
    - **Task 2: Classify fire causes (multi-class classification).**
        - Train a LightGBM model to predict fire causes (e.g., "lightning," "industry").

        - Assess performance with precision, recall, and F1-score.
    - **Task 3: Predict fire size class (multi-class classification).**
        - Use Random Forest to predict fire size classes (e.g., small, medium, large).
        - Evaluate using accuracy, precision, recall, and F1-score.
- **Visualization:**
    - Create GIFs to visualize predictions over time, grouped by year.

Results

Task 1: Wildfire Occurrence Prediction

- Random Forest achieved:
    - Accuracy: 100% (illogical, needs to be check for data leakage and overfitting)
- Early detection of fire-prone conditions enhances readiness.

Task 2: Fire Cause Classification

- LightGBM achieved:
    - Accuracy: 64% across all fire causes.

Task 3: Fire Size Class Prediction

- **Random Forest achieved:**
    - Accuracy: 94%.
    - Higher precision for larger size classes, indicating robust predictions for severe fires.

Key Insights:

- Rolling averages and temporal features significantly improve predictions.
- Handling class imbalances is crucial for rare fire causes and larger size classes.

Future Work

Immediate Improvements:

- Checking for data leakages and overfitting, mainly the fire occurrence prediction.

- Slope and wind speed integration for fire spread and its direction prediction.

Long-Term Goals:

- Incorporating satellite data and live sensors (heat anomalies, vegetation dryness, and smoke plumes)
- Developing evacuation plans using population and infrastructure data.
- Integrating climate change projections into models.

Guide and Description:

The files described below are the notebooks used in the project, they can be found in the **All codes folder.**

1. Final Model: Uses the Random forest model to Predict wildfire occurrence, wildfire causes, Includes hyperparameter optimization using GridSearchCV and dataset balancing via SMOTE. LightGBM Class is used as an alternative to enhance performance and speed. Visualizations include Heatmaps and scatter plots for wildfire data on Alberta maps and Animated GIFs illustration of yearly trends in fire occurrence and causes. **This is the main file that has the entire code and most accurate model. Run this.**
2. EDA: General Exploratory Data Analysis for understanding the causes and location of the fires and conducting PCA.
3. XG_Boost: This file focuses on predicting wildfire causes and impacts using XGBoost, with steps for preprocessing, training, evaluating models, and visualizing fire data on Alberta's map. This is another model we tried, to compare its accuracy to other models and see which one is more accurate and computationally efficient
4. NN: This file processes wildfire and weather datasets, performs data cleaning and feature encoding, builds Random Forest and neural network models to predict wildfire attributes (e.g., size and spread rate), and visualizes results on Alberta's map. This file was used as a comparison between the results of the different models to see which one is more accurate and saves computational resources.
5. last version: Has similar models as the Final Model but shows that there are no common latitude and longitude values found between df1, df3 merged file and any of the groups in df2.
6. RF: First version of Random Forest tried (not well optimised).

All the datasets used can be found in the datasets folders.

References:

[1] https://www.ratehub.ca/blog/canadian-wildfires-insurance-industry/

[2]https://natural-resources.canada.ca/climate-change/climate-change-impacts-forests/forest-change-indicators/cost-fire-protection/17783

[3] https://wildfire.oregon.gov/evacuations

[4] https://www.alltrails.com/canada/alberta

[5] https://simplemaps.com/static/data/canada-cities/1.8/basic/simplemaps_canadacities_basicv1.8.zip