

Project Proposal

4AL3 Applications of Machine Learning - Group 50

Eric Solak

Tyler Yue

Ahmed Sahi

Overview

Task Title: Predicting Wildfire Risk Using Environmental Data

Overview

Wildfires are a growing threat to infrastructure, ecosystems, and human safety. Predicting the likelihood of wildfires in advance can help authorities, first responders, and communities with preparation, resource allocation, and reduce potential damage. This task focuses on developing a machine learning model to classify the wildfire risk level of a given region using environmental and meteorological data.

Significance and Challenges

Wildfires are increasing in frequency, intensity, and scale, and wildfire activity has doubled over the last 21 years [1]. This is due to climate change which has created longer fire seasons, drier conditions due to reduced precipitation, and warmer nighttime temperatures. The risk of wildfires extends beyond the destruction of land and property. Wildfire smoke can travel thousands of kilometres, ruining air quality and putting millions of people at risk [1]. This model's risk classification can reduce the risk for millions of people by providing insights to help first-responders allocate firefighting resources ahead of time.

There are several challenges that make this difficult:

- The relationship between environmental data and wildfire risk is non-linear [2].
- The same weather conditions may have different risk levels in different regions due to differences in terrain and vegetation [1].
- Fires can be ignited by human activity which can lead to outliers or noise in data.
- Wildfires are more rare to occur than regular conditions which can bias the model to predict “low risk” in general. This can make it harder to accurately recognize “high-risk” conditions.

Task Definition

Data Type

The type of data which the model will be trained on is environmental, positional and meteorological data with numerical features.

Task Type

This is a classification task. The model will predict the wildfire risk level of a given region.

Number of Classes

There are 3 classes for the risk level: low-risk, medium-risk, high-risk.

Label Type

The model will perform single-label classification, assigning one label per instance.

Data

For this project, we will utilize a Wildfire Prediction Dataset found on [Kaggle](#). This dataset contains 118858 entries and 17 attributes, with no missing values. Further, all values are numerical (float64), simplifying any data manipulation by avoiding issues revolving around categorical data. There are two main sources used by this dataset: NASA's FIRMS VIIRS SCC system for fire radiative power (FRP) data, and Open-Meteo's meteorological data. As these sources are open and publicly available, the data used is compliant with all terms of service. Below is a sample of 3 entries from the dataset:

daynight_N	lat	lon	fire_weather_index	pressure_mean	wind_direction_mean
0	-15.1993	38.54393	5.654271	955.6083	136.0833
1	31.51203	-101.575	16.56467	927.0167	181.3333
1	-13.7454	28.05493	5.542089	884.3792	109.1667

wind_direction_s td	solar_radiation_me an	dewpoint_me an	cloud_cover_me n	evapotranspiration_to tal	humidity_m in
43.61185	250.3333	15.88333	18.375	4.89	35
31.92526	296.9167	10.52917	3.416667	7.24	16
8.37987	210.9583	6.095833	12.20833	3.68	35

temp_mean	temp_range	wind_speed_max	occured	frp
23.175	12.9	13.5	0	3.69
29.73333	13.4	16.2	0	0.73
14.10417	11.1	16.8	0	0.78

As this project aims to use the dataset to predict wildlife risk (low, medium, high), this is a classification problem with three categories, for which the labels for each entry will be determined using the 'fire weather index' and the 'occured' variable. For the low-risk category, entries with a low fire weather index, and no recorded fire event will be flagged. Index values and fire-conductive weather

patterns which fall towards the median will be labelled as medium-risk, with the exact thresholds currently undecided. Lastly, high-risk entries will have fire weather index values towards the upper end, or a flagged fire occurrence. Therefore, the preexisting data will be used to create this new label field, without any manual intervention, apart from determining the mentioned thresholds. A standard training/validation/testing split will be used, e.g. 70/15/15, and the non-binary fields will be normalized for better performance.

High-level Solution Description

In order to go about solving this problem we plan on employing a supervised machine learning classification model that will classify a specific region as a low, medium or high wildfire risk area. This model will take in our dataset of features 17 numerical features. The features include positional, meteorological and environmental data. These features describe the environmental conditions that influence fire behavior and ignition potential. The model's target variable will be the wildfire risk category, derived from the fire weather index and occurred variables in the dataset. Since the relationship between environmental variables and wildfire risk is highly non-linear, we plan to use a strategy that can classify non linearly separable data. This will include using feed forward neural networks to learn the complex interactions between the positional environmental and meteorological data. Using activation functions such as ReLU or sigmoid will allow the model to learn the non linearly separable data, and a softmax function to convert the outputs into probabilities. Additionally, the model will require regularization to prevent overfitting the data. Existing work in this space has two main focuses. First there are models that predict the spread of wildfires based on satellite imagery of the wildfire along with other data [3]. These models use Deep neural networks to accurately predict the fire spread. Secondly, there are models that predict the likelihood of wildfires rather than their spread, which aligns closely with our solution. Current techniques include XGboost for linear regression and classification and statistical models including latent Gaussian models and integrated nested Laplace transforms [4] to forecast wildfires. We believe that these models can be improved with the use of neural networks. Canada currently employs a system called Canadian Forest Fire Danger Rating System (CFFDRS) which is a deterministic model that uses empirical data from experiments to determine wildfire risk, this model employs logistical regression. Model performance will be evaluated using a combination of accuracy, precision, recall, and F1-score. We will also analyze a confusion matrix to identify which risk categories the model struggles to distinguish. We will implement our models Pandas and NumPy for data preprocessing, Scikit-learn for model evaluation metrics, and PyTorch for creating our neural network.

Works Cited

- [1] NASA. (2025, May 28). *Wildfires and climate change - nasa science*. NASA.
<https://science.nasa.gov/earth/explore/wildfires-and-climate-change/>
- [2] Hydrometeorology-wildfire relationship analysis based on a wildfire bivariate probabilistic framework in different ecoregions of the continental United States - sciencedirect. (2024, September 3). <https://www.sciencedirect.com/science/article/pii/S0168192324003289>
- [3] *Predicting Next-Day Wildfire Spread with Time Series and Attention*. (2024). Retrieved October 8, 2025, from <https://arxiv.org/html/2502.12003v1>
- [4] Hu, C., Bispo, R. B., Rue, H., DaCamara, C. C., Swallow, B., & Castro-Camilo, D. (2025). *XGBoost meets INLA: a two-stage spatio-temporal forecasting of wildfires in Portugal*.
<http://arxiv.org/abs/2508.09896>