Abstract

Wildfires pose significant challenges to ecosystems and human communities, necessitating improved prediction and mitigation strategies. This document presents a comprehensive study that utilises diverse datasets and Machine Learning (ML) models to enhance wildfire risk and severity predictability in Alberta, Canada, from 2006 to 2021. By integrating historical wildfire data, ERA5 weather conditions, MODIS satellite vegetation indices, demographic information, and additional variables such as proximity to roads and population density, our research aims to identify high-risk zones and vulnerable regions effectively. We developed two ML models: one to predict fire ignition probability based on weather and vegetation indexes and another to estimate the severity of wildfires by analysing burned area polygons. This dual-model approach enables early detection of potential danger zones and assessment of fire impacts. Our study also includes statistical analysis to identify patterns in wildfire causation, contributing valuable insights into effective prevention and mitigation strategies. Despite assumptions about the reliability of historical data and model accuracy, the potential value of our work in improving wildfire management strategies is significant, offering a promising direction for future research and innovation in wildfire prediction and management.

Introduction

This study addresses the urgent need for advanced wildfire prediction and mitigation strategies through a detailed analysis using diverse datasets and Machine Learning (ML) models. By integrating historical wildfire data, weather conditions, vegetation indices, and demographic information from Alberta (2006-2021), we aim to enhance the predictability of fire risk and severity and identify vulnerable regions.

Background. Leveraging a wide range of data, including ERA5 weather data and MODIS satellite observations, our research utilises advanced data management and preprocessing techniques to ensure a robust analytical foundation.

Research and Applications. Our research develops two ML models to predict fire ignition probability and estimate fire severity. These models are designed to work in tandem, enabling early detection of high-risk areas and assessing potential wildfire impacts. The inclusion of statistical analysis reveals patterns in wildfire causation, offering insights into effective mitigation strategies.

Assumptions. The study assumes the reliability of historical data to forecast future wildfire trends despite efforts to correct data inaccuracies. It also relies on the assumption that our models can accurately reflect the complex dynamics of wildfire behaviour.

Value. The primary value of our work lies in its potential to improve wildfire management and response strategies. By providing predictions of wildfire risks and impacts, our research supports more effective resource allocation, preventative measures, and mitigation efforts, ultimately contributing to the safety and well-being of communities in wildfire-prone areas.

Methodology

In this section, we start with the fundamental analysis: explain what and why data we used, which Machine Learning (ML) models were used, and finish with statistical analysis to answer EY challenge questions.

Used data. We weren’t limited to the given data. Additionally, we downloaded weather and forest indexes for predicting fire risk and severity and population data to highlight vulnerable regions.

- Historical wildfire data of Alberta (2006-2021) [1].
- ERA5 [2] weather data (2006-2021): the eastward 10m wind (u), the northward 10m wind (v), total precipitation, and dry-bulb (T) and dew-point (Td) temperatures. 12 PM timeframe.
- MODIS [3] satellite data (2006-2021): NDVI (vegetation density), LAI (Leaf Area Index).
- Historical burned area polygons of Alberta (2006-2021) [4].
- Static variables: Distance to roads and population density (2015) [5, 6].
- Indian Reserve and Métis Settlement [7] (2024).

Data load. The open-source library OpenDataCube[8] stores and reprojects all obtained variables. All parameters were reprojected to EPSG:4326. An example of Era5 weather data:

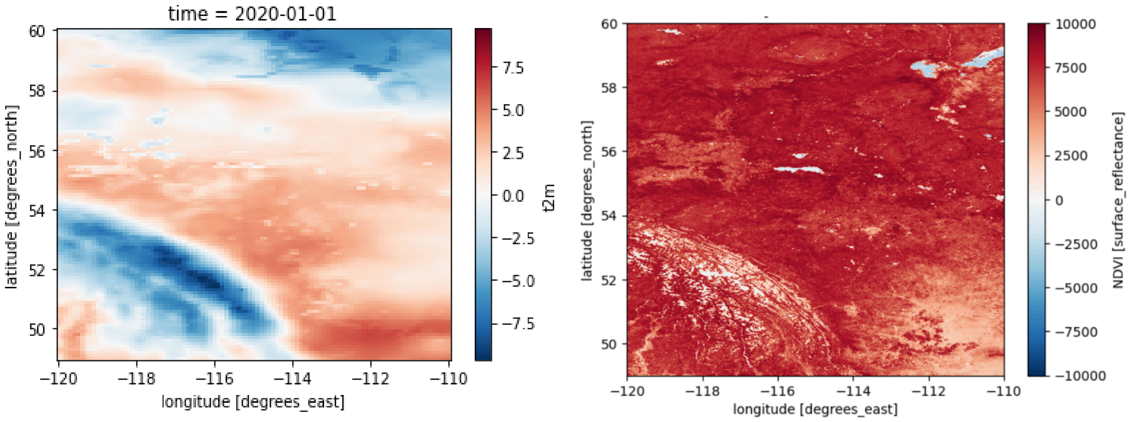


Figure 1 Era5 temperature and MODIS NDVI for Alberta

Data preprocess. We conducted data preprocessing to remove data misinformation, interpolate data and perform calculations.

- Historical wildfire data: several datetime (dt) columns didn’t seem to have reasonable data. For example, `fire_start_date` contains wrong years: at index #0, the year 2010 was replaced with the year 2020 (as other dt columns described the year 2020); #1291 - replaced 1021 with 2021; and #14316 - 0201 with 2011. While the most apparent wrong years were fixed, we still had two cases: NaT and a huge (more than 80 days) time gap between the fire start and the assessment date. To handle this, we subtracted `assessment_date` from `fire_start_date` and took a 98 percentile (=80 days difference) as a condition to use `assessment_date` as `fire_start_date` (same for NaT data). The decision to use `assessment_date`, and not other “earlier stages” columns, was based on two factors: 1) `assessment_date` has no absent data and points on assessed fires (severe or potentially severe), and 2) the difference between assessment and other stages is short and reasonable. The rest of the columns remained unprocessed.
- ERA5: relative humidity (RH) parameter was calculated using the following formula [6]:

$$RH = 100 * exp(17.625 \times 243.04 \times \frac{Td - T}{(243.04 + T - 273.15) \times (243.04 + Td - 273.15)})$$

Wind direction (1) and speed (2) were calculated like [9]:

(1) $$\phi = mod(180 + \frac{180}{\pi} atan2(v, u), 360)$$

(2) $$|\overline{V}| = \sqrt{v^2 + u^2}$$

- MODIS: remote sensing data is observed every 16th day for NDVI and every 8th day for LAI. Linear interpolation was used to fill in time gaps.

AI Models. We developed two models for two tasks:

1. Estimating fire burst probability - highlighting potential dangerous zones depending on weather and vegetation indexes.
2. Predict burned area for appeared fires - estimate severity for surrounding areas.

Now, more details per model:

1. All weather parameters with both MODIS indices were used for this task. Also, the static variables were included: population density and distance to roads. For every time T when the wildfire occurs, the variables for the previous 32 days([T-32, T]) were taken to predict wildfire probability. Years 2018 and 2019 were taken for validation, 2020 and 2021 for testing, and all remaining for training. All data were resampled to the same resolution - 500m. Standardisation was done to set values in the same range. Negative samples were also generated to predict probability. Negative samples (random 32 days guaranteed not to cause wildfire) were generated two times more than positives as in [10]. Negative samples were distributed through the years the same as the positive samples. Two-layer LSTM model(128, 128) with 3 FC layers(128, 64, 32) was used to output the probability of wildfire. From this, the baseline was created with an F1 score of 0.6153. The F1 score is used here to not only set high probabilities for potential wildfire locations but also not to mark safe locations as dangerous. To improve the classifier, we used the `curren_size` field from the dataset - total burned size. Adopting the idea from [11], we calculated the weighted binary cross entropy loss where weights are the values of the total burned area. Because this data has a significant deviation(from 0.01 to 4000000), log10 was taken to make smaller weights with smaller deviations. This approach allowed us to improve the F1 score to 0.689.

The following images represent the 2020 year:

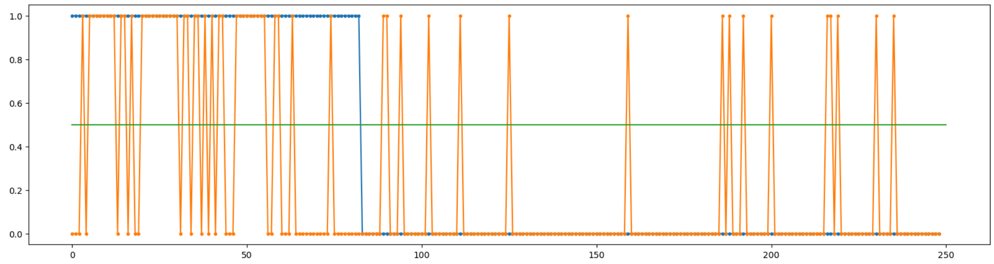


Figure 2. Fire vs no fire 2020 baseline. Blue line - ground truth. Green line - decision line (0.5 threshold). Orange line - predicted fire danger

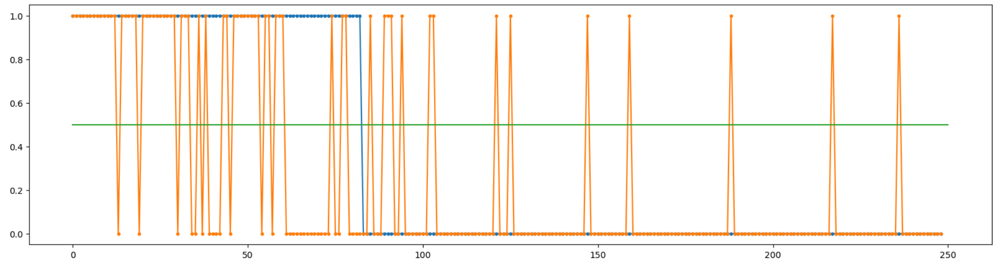


Figure 3. Fire vs no fire 2020 improved. Blue line - ground truth. Green line - decision line. Orange line - predicted fire danger

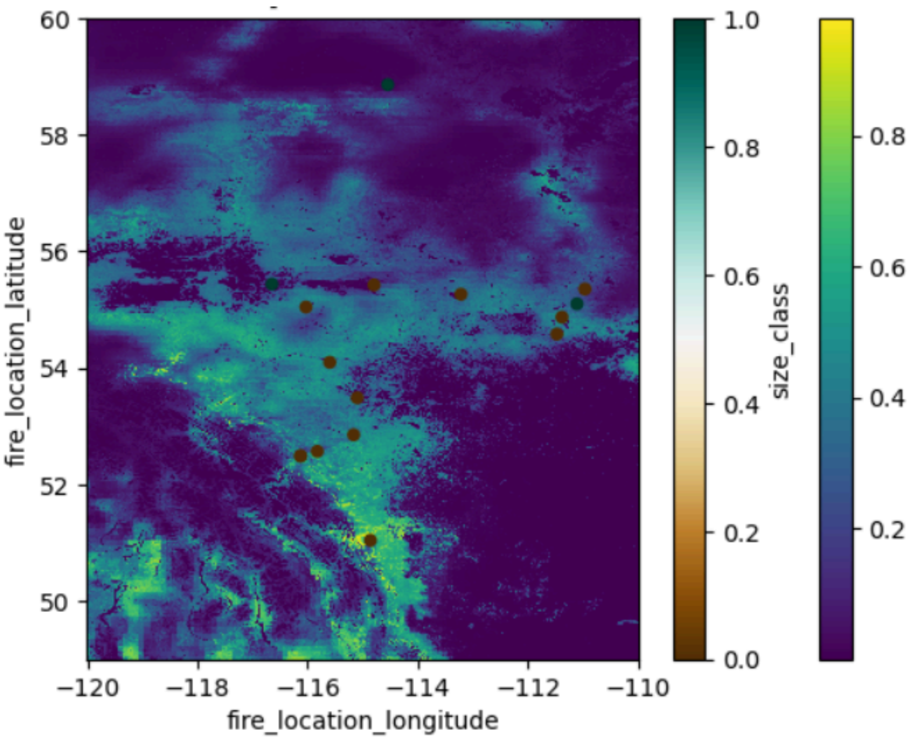


Figure 4. Prediction for the mid-August 2021. The brighter pixels correspond to the more dangerous regions

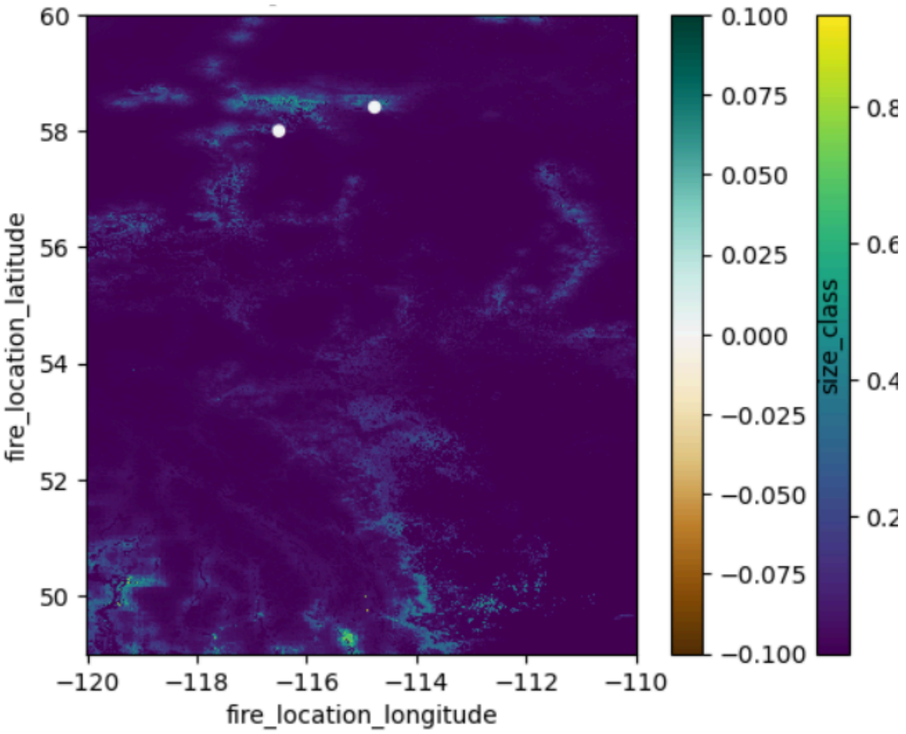


Figure 5. Prediction for the beginning of September 2021. The brighter pixels correspond to the more hazardous regions

2. The main objective of this task is to predict the burned area from some ignition point from the beginning of the wildfire. We used a dataset with polygons that show what area was burned to predict the burned area. Not all polygons are mapped to the initial data. So, only polygons that matched the current size, time, and location were taken. If the ignition point is close enough(less than 40% of the length of the polygon), the new ignition point is chosen - the nearest point from the real ignition and polygon(Fig. 6). The area of 64x64 pixels (32x32 km) taken around the ignition point. All polygons could fit in this region. For model training, the 32x32 pixels patches were cut from the main tile, and this brought the data augmentation.

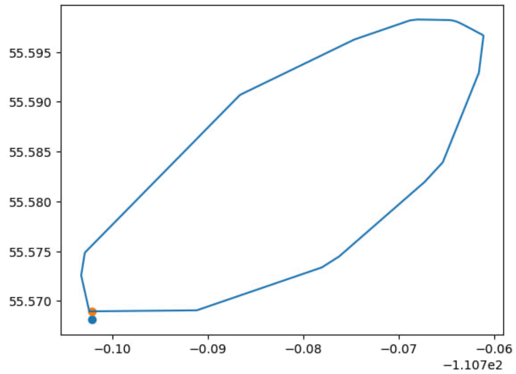


Figure 6. The burned area polygon with original ignition point(blue dot) and new ignition point(orange dot)

The same data was used to predict the final shape of the burned area, but also the ignition point mask was also added. The U-net model with ResNet backbone was used to make the binary segmentation of burned pixels. The resulting area is the number of pixels. Because the spatial resolution is 500m, the output area is not very precise; however, for big fires, knowing how and how fast fire spreads is acceptable.

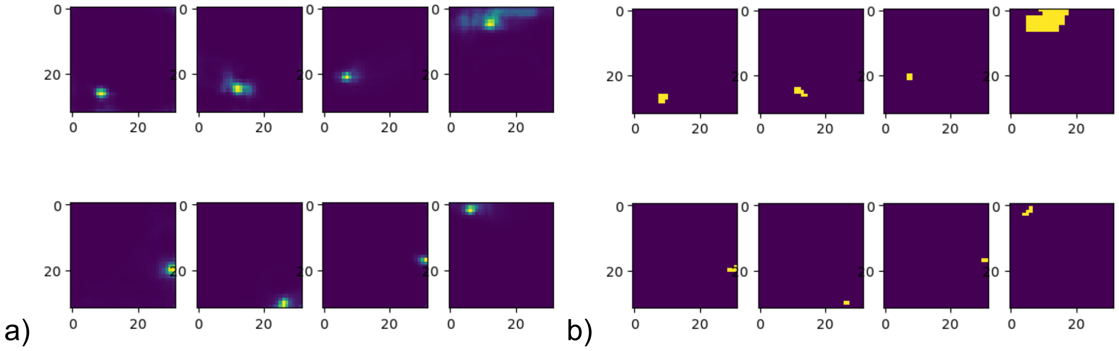


Figure 7. Burned area model results: ground truth (a), model output (b)

The primary purpose of both these models is to work in a pair. As soon as the first model detects the dangerous region, the second one can predict the size of the wildfire and warn residents.

Statistical research. Human-caused Alberta wildfires are the absolute majority (Fig. 8a). In contrast, lightning-caused wildfires are the absolute majority in the total burned area (Fig. 8b).

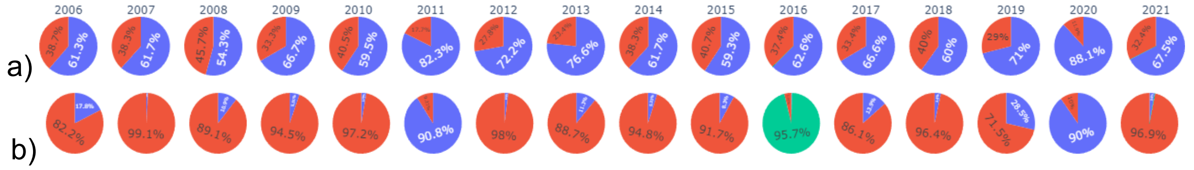


Figure 8. Percentages of a whole number of fires (a) and burned hectares per cause (b). Blue - human-caused; red - lightning; green - under investigation.

This explains that human-caused fires rarely burst to the same levels compared to lightning cases. The top 3 reasons for human-caused wildfires are cooking and warming (unsafe fire) - 4701.67 Ha total from 2006 to 2021, off-road vehicles (burning substance) - 2560.56 Ha, and burning without a permit (grass) - 2022.08 Ha.

Objectives & Results

1. **Which top FSA regions are more vulnerable to Wildfires? Assess the vulnerability of the population in each vulnerable FSA region and identify the wildfires' impact on the indigenous population.**

To answer this question, we used the output of the first AI model from above.

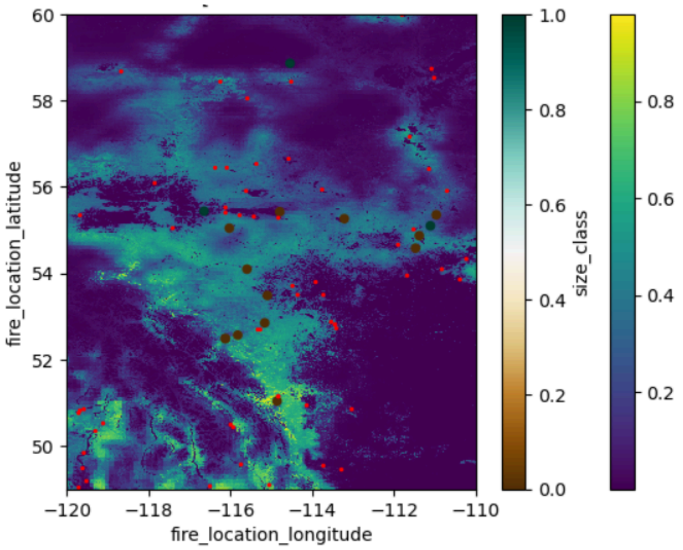


Figure 9. Indian Reserve on top of the vulnerable area predictions for mid-Aug 2021.

Below we provide a map of fire distribution based on main human causes.

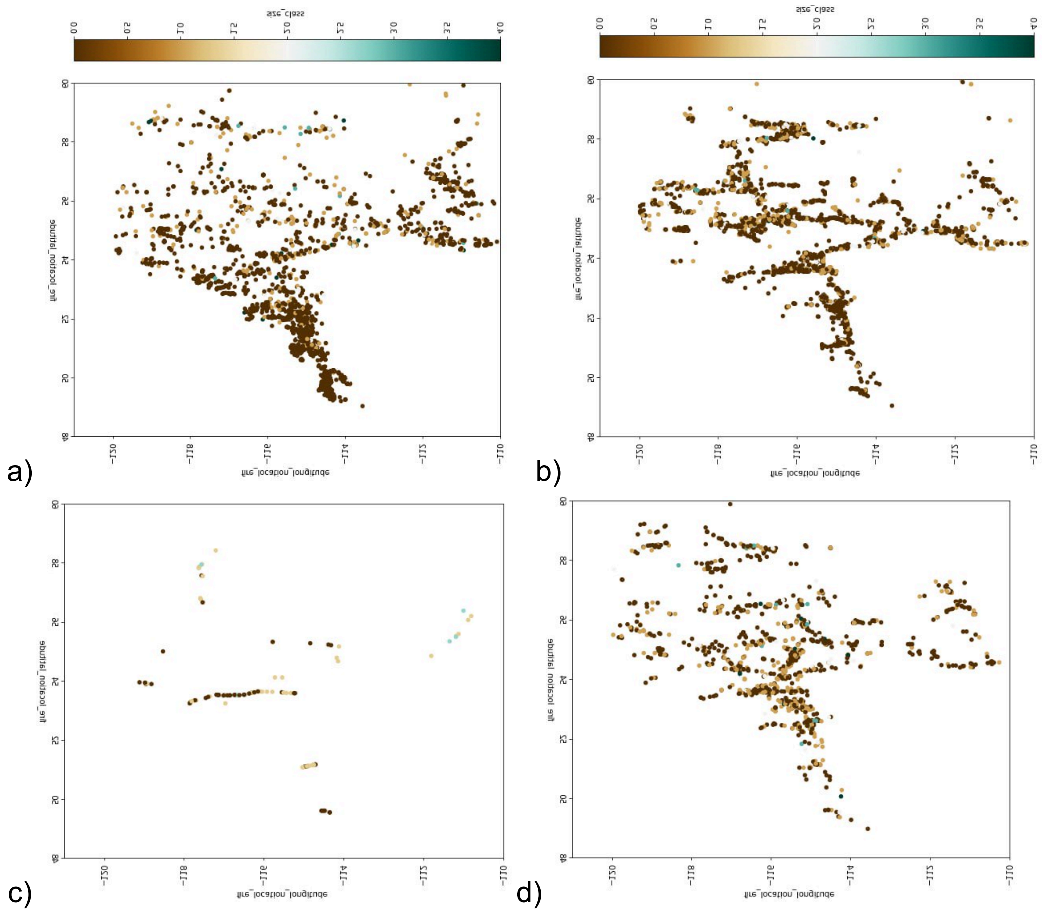


Figure 10. Recreation (a), resident (b), railroad (c), power line industry (d)

Based on these maps, we can see that tourists and locals impact on Indian Reserves areas. Many fires are caused in popular tourist places, and nothing less than locals cause many fires too by debris and refuse disposal and traditional burning.

2. **What are the main reasons that usually cause the wildfires near each vulnerable FSA region? And especially analyse the main reasons for the wildfires with large burn areas and much bigger impact on the environment and residents close by.**

We included vulnerability factors dependency in our AI model: weather conditions, vegetation (forest type), and month of the year (seasonality).

We found out that statistically large fires appear during the summer months (May-Jul) (Fig. 11a).

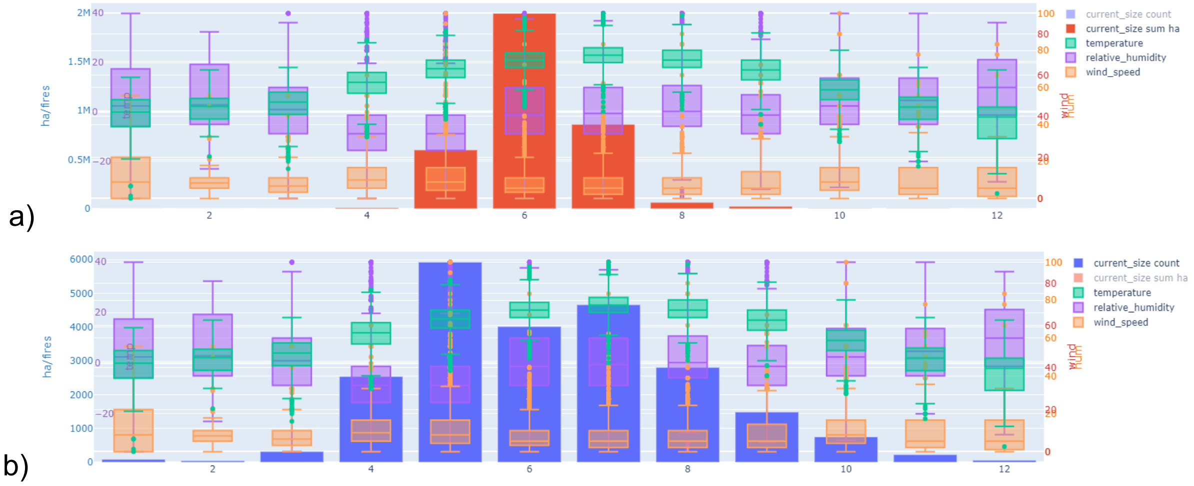
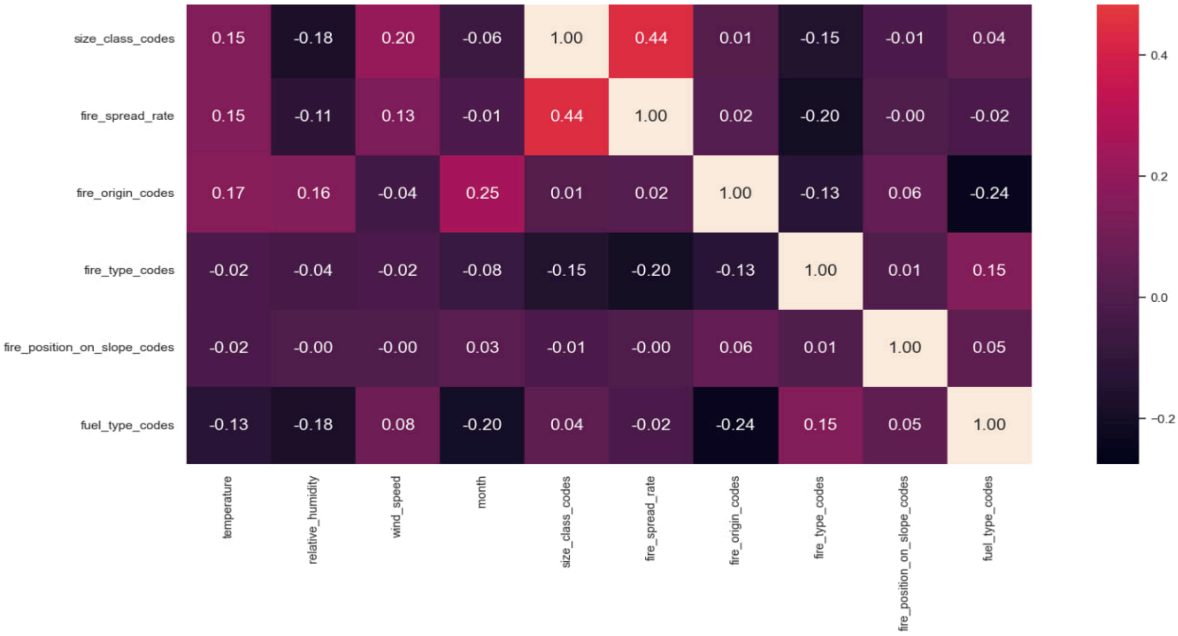


Figure 11. Red bars - the sum ha of burned area (a), blue - number of fires (b); box plots are - the impact of weather on wildfires.

Additionally, here is a correlation research:



	temperature	relative_humidity	wind_speed	month	size_class_codes	fire_spread_rate	fire_origin_codes	fire_type_codes	fire_position_on_slope_codes	fuel_type_codes
size_class_codes	0.15	-0.18	0.20	-0.06	1.00	0.44	0.01	-0.15	-0.01	0.04
fire_spread_rate	0.15	-0.11	0.13	-0.01	0.44	1.00	0.02	-0.20	-0.00	-0.02
fire_origin_codes	0.17	0.16	-0.04	0.25	0.01	0.02	1.00	-0.13	0.06	-0.24
fire_type_codes	-0.02	-0.04	-0.02	-0.08	-0.15	-0.20	-0.13	1.00	0.01	0.15
fire_position_on_slope_codes	-0.02	-0.00	-0.00	0.03	-0.01	-0.00	0.06	0.01	1.00	0.05
fuel_type_codes	-0.13	-0.18	0.08	-0.20	0.04	-0.02	-0.24	0.15	0.05	1.00

Figure 12. Fire class (size) depends on the weather, spread rate, and type of fire behaviour (crown/ground). The spread rate depends on behaviour and weather. Fire behaviour depends on the soil (fire fuel)

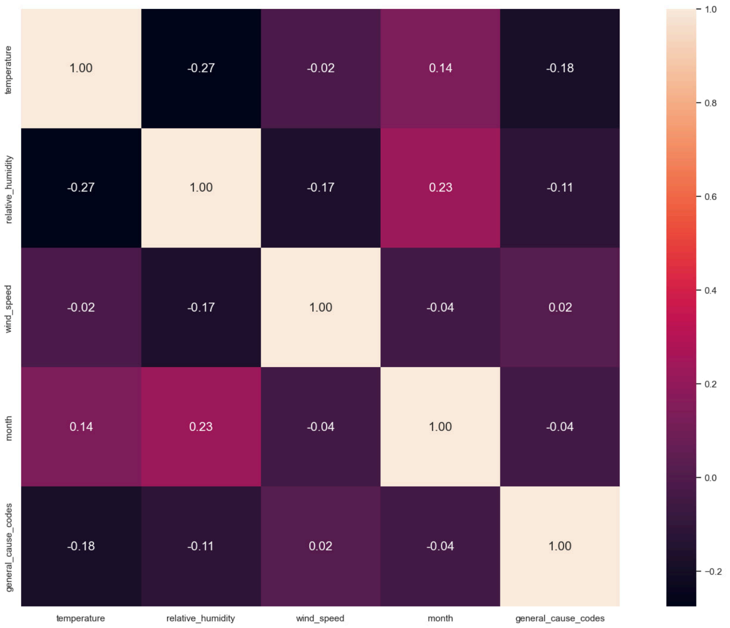


Figure 13. Causes depend on weather parameters (temperature and humidity)

Future work. Add weather station data, calculate the FWI index [13], and add carbon consideration. Models improvements.

Conclusions

Our study combines extensive data analysis and Machine Learning models to enhance wildfire prediction and management strategies. By integrating diverse datasets on weather, vegetation, and demographics with advanced ML techniques, we have developed predictive models that identify high-risk zones and estimate wildfire severity. This approach not only addresses the EY challenge but also provides valuable insights into wildfire dynamics, contributing significantly to the development of targeted mitigation and prevention strategies. Our findings underscore the potential of data-driven models in improving the effectiveness of wildfire management efforts, highlighting the importance of continued research and innovation in this critical field.

References

1. https://www.alberta.ca/wildfire-maps-and-data
2. Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. (2023): ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.adbb2d47 (Accessed on 21-Feb-2024)
3. https://modis.gsfc.nasa.gov/data/
4. https://cwfis.cfs.nrcan.gc.ca/datamart
5. https://hub.worldpop.org/geodata/listing?id=64
6. https://hub.worldpop.org/geodata/listing?id=33
7. boundary/first_nations_land (MapServer) (alberta.ca)
8. https://www.opendatacube.org/
9. https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022GL099740

10. Spyros Kondylatos, Ioannis Prapas, Michele Ronco, Ioannis Papoutsis, Gustau CampsValls, María Piles, Miguel-Ángel Fernández-Torres, and Nuno Carvalhais. Wildfire Danger Prediction and Understanding With Deep Learning. Geophysical Research Letters, 49(17):e2022GL099368, 2022. ISSN 1944-8007. doi: 10.1029/2022GL099368.
11. Kondylatos, S., Prapas, I., & Papoutsis, I. (2023). Mesogeos: A multi-purpose dataset for data-driven wildfire modeling in the Mediterranean. *ArXiv*. /abs/2306.05144
12. https://confluence.ecmwf.int/pages/viewpage.action?pageId=133262398
13. https://www.frames.gov/documents/alaska/c5/files/FWI-history.pdf