BCIT SMART

BC wildfire predictive model

Instructor: Linda Butterfield

Tolulope Adegboye, Dien Vo, Chi-Yu Lee
2025/05

Table of Contents

Executive Summary

Wildfires occur in British Columbia every year, and in recent years, the affected areas have been expanding. If their occurrence can be predicted in advance, it may be possible to effectively reduce damage and losses. In this report, we present historical trends of wildfires and develop a predictive model for wildfire occurrences based on geographic and meteorological information. We take two approaches towards building the model: a geospatial approach, and a Python-based approach.

From analysing the data, we determined some key findings:

From 1990 to 2022, wildfire activity was generally observed in the second and third quarters of the year. Recently, there has been a trend of increasing wildfire impact.

Through both modeling approaches, it was discovered that the random forest model was the most accurate. Out of all the features observed, for the geospatial model, the temperature (dewpoint and 2-metre temperature) and precipitation were the 3 most important features in order. However, for the Python model, wind speed was the 3rd most important instead of precipitation.

The geospatial random forest model had a high sensitivity, but a lower precision and accuracy compared to the Python model. This highlights the increased ability to detect fires, but its potential to predict certain fires that are not present.

From both the XGBoost and random forest model, increasing the 2-metre temperature corresponds to increases in the fire probability. However, in the random forest model, increases in the 2-metre dewpoint temperature and total precipitation generally are associated with decreases in the fire probability.

From the correlation matrix, we saw a picture of how different variables relate to property damage. Based on the results, the 2-metre temperature shows the highest correlation with property loss.

The impact of temperature can be further highlighted in the seasonal risk prediction as a significant temperature increase corresponded with an increase in the number of fire points from July to August.

Recommendations

BCIT SMART can use these models to make forecasts based on the initial data points and then monitor certain areas that have a higher risk of fires. It is important to monitor 2-meter temperature, 2-metre dewpoint temperature, and total precipitation values daily to detect areas with higher risks of fires.

Benefits

Through this, wildfire services and local utilities are able to adjust their operations and implement different strategies proactively to lower the impact of wildfires on the environment and consequently the property damages.

Introduction

The goal of the project is to create a preliminary model to predict wildfire risk that can be further expanded and improved by other researchers. Such a model can help mitigate wildfire risk in the future. Authorities can potentially use these predictive models to readjust their spending to avoid property loss and damage from the fires, plan resources in advance, and locate equipment appropriately to respond quickly.

In recent decades, wildfires have caused widespread environmental damage, economic loss, and displacement of communities across many regions, including British Columbia. To prevent and mitigate the impacts of wildfires,

many valuable studies have been conducted in the past. In this project, we focus specifically on wildfires in British Columbia, Canada. The goal of this project is to develop a predictive model that estimates wildfire risk using historical data. By analyzing patterns in past wildfires, weather conditions, and other environmental factors, the model aims to identify areas at higher risk.

Previous studies have explored various approaches to wildfire prediction and risk assessment. Regarding the predictive model, Traditional model solely based on Fire Weather Index, Wagner (1974). However, McNorton (2024) developed a predictive model using a vegetation characteristic model, weather forecasts, and a data-driven machine learning approach. Mohajane et al. (2021) developed five new hybrid machine learning algorithms.

Items within the scope of this project include:

- Sourcing data
- ETL and visuals
- Providing the descriptive analysis of the data
- Observing the relationship between area burned, temperature and property damage in dollars
- Creating a model to predict the probability of fires in BC
- Providing detailed reports and giving a (mid-point & final) presentation to BCIT SMART

Items outside the scope include:

- Gas emissions analysis
- BC towns data analysis
- Formatting the files for seamless technical integration into other systems
- Insurance premiums predictions

Methodology

We decided to take two approaches to predicting wildfires: a geospatial approach, and a primarily Python-driven approach.

Data Sources

We used Canada's National Forestry Database to analyze historical trends on burned areas and fire occurrences. It is relevant to provide an appropriate overview of the situation in BC. More information can be found in the data sources section of the appendix.

We also obtained fire data from National Resources Canada from 2000 to 2024 in the form of hotspots. Each hotspot data file contains information on where and when each fire occurred. It provided a starting point to understanding fire characteristics. More information can be found in the Data Sources section of the appendix.

From the same website, we obtained data on the BC area fuel type characteristics that were classified into different types such as spruce-lichen woodland, or boreal black and white spruce. As a result, we were able to identify the conditions of which a fire occurred to further understand which types are more susceptible to fires and which ones are not as susceptible. One assumption made was that the fuel types in BC remained the same on a yearly basis.

For the second data-building approach, we were inspired by the article A Global Probability-Of-Fire (PoF) Forecast by J.R. McNorton, published in June 2024. The study provided valuable insights that helped us finalize the selection of variables for our project. It recommended expanding the variable set by including factors such as fuel type, high leaf area index, and dew point temperature—key elements derived from the traditional Fire Weather Index used in fire risk prediction. According to the article, incorporating these additional features significantly improved the model's predictive power, stating that it "outperforms existing indices, providing accurate forecasts of fire activity up to 10 days in advance, and in some cases up to 30 days." Although we did not follow the exact 9 km resolution method used in the study, we adopted its general approach to explore how including these variables would affect our model accuracy and the overall quality of the resulting database.

With that in mind, we used two different data sources recommended in the

article. For fire occurrence data, we used the **MODIS Fire** dataset, which includes historical records of where fires actually took place. For climate data, we used the **ERA5 monthly data on a single level**, which provides a wide range of climate variables from 1940 to the present. There are a few limitations in the dataset, particularly concerning variables such as temperature and dew point temperature. These limitations are further illustrated in the appendix below.

The MODIS dataset captures real fire events, while the ERA5 dataset includes other important environmental factors such as temperature, dew point, wind speed, and more. By combining these two sources, we aimed to create a more complete dataset for modeling fire risk.

Geospatial Approach

Most of the data cleaning was conducted in PyGIS, a coding platform that enables Python coding within QGIS, a geographic information software. Through the platform, the locations of each BC fire hotspot were extracted, then grouped by month and year. These points were given a 'Fire' label of 1. Random generated points were generated to represent locations without fires which were classified by the 'Fire' label of 0. More information on the point generation process and limitations can be found in the appendix.

With the points, along with the ERA5 monthly climate data and fuel type layer, the relevant information was able to be sampled. Each point had the exact temperatures, wind data, precipitation, leaf area index, and fuel type based on its location. This information was then exported to a CSV file in Python for predictive modeling and analysis.

Python Analytics

The goal is to merge three datasets: the MODIS fire occurrence data, which records all past fire events; the ERA5 monthly climate data, which provides historical weather conditions; and the fuel type dataset, which includes classifications for different fuel types. We first created a column called **Fire Occurrence**, assigning a value of **1** to all records from the MODIS fire dataset. This dataset was then merged with the climate data based on **latitude,**

longitude, month, and year, allowing us to assign a value of **0** to records from the climate data that did not match any fire event. Finally, we completed the database by incorporating fuel type information using spatial matching coordinates between the merged data and the fuel type dataset. A detailed demonstration of this process, along with its limitations, is provided in the appendix below.

Variables

Variables	Description	Why is it useful?
T2m – 2-metre temperature (K)	The air temperature is 2 meters above ground level	High temperatures dry out vegetation, making it more flammable and increasing fire risk
d2m – dewpoint temperature (K)	Temperature at which air becomes saturated with moisture (dewpoint)	Lower dewpoints indicate drier air, which helps vegetation dry out and become more ignitable
10 U & 10 V wind (m)	East-west and north-south wind components measured at 10 meters height	Wind drives fire to spread, oxygen supply, and can carry embers over long distances
Tp - Total Precipitation (m)	Total rainfall	Lack of rain leads to dry conditions and fuel buildup, increasing fire potential
Fuel type	The type of vegetation or material that can burn on the ground	Different fuels ignite and burn at different rates, affecting how a fire spread
Leaf Area Index (lai_hv)	Amount of leaf coverage per unit area, representing vegetation	Higher LAI means more available fuel and more intense, longer-lasting

	density	fires.

Analysis

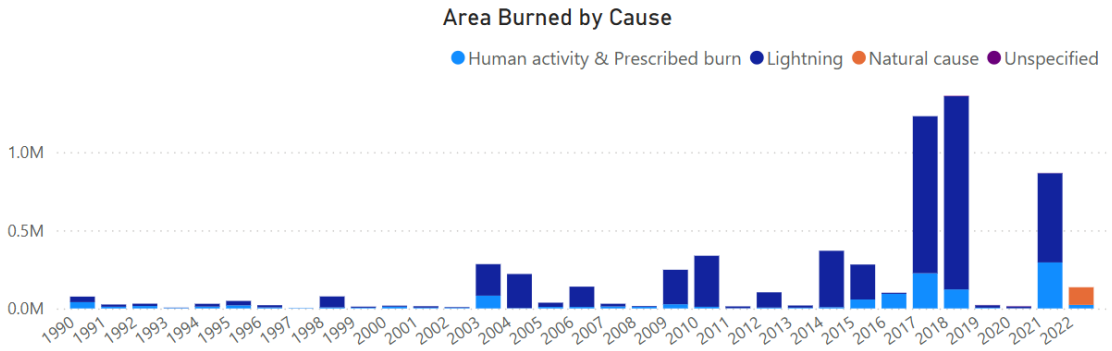
We begin by examining historical trends in wildfire activity to gain insights into past patterns of burned area, causes of wildfires, and associated property loss.

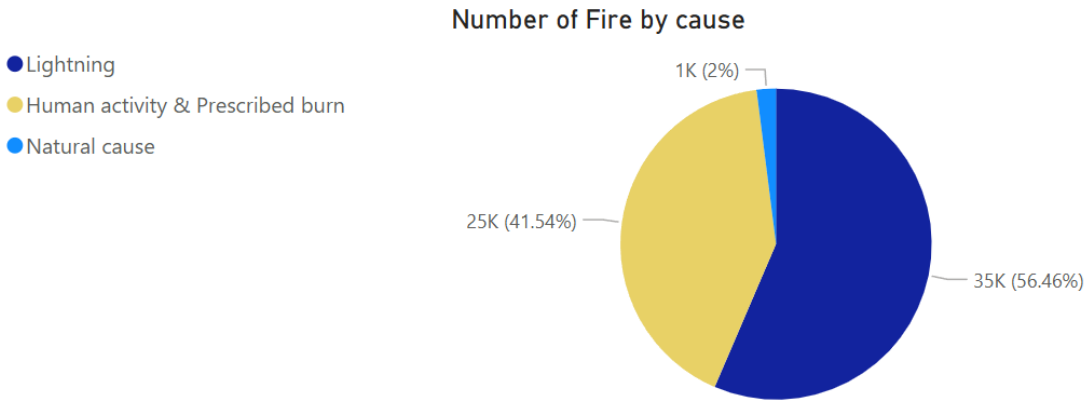
From 1990 to 2022, burned areas typically reach its annual peak in July. Overall, wildfire activity is concentrated in the second and third quarters of the year.

Top 5 months

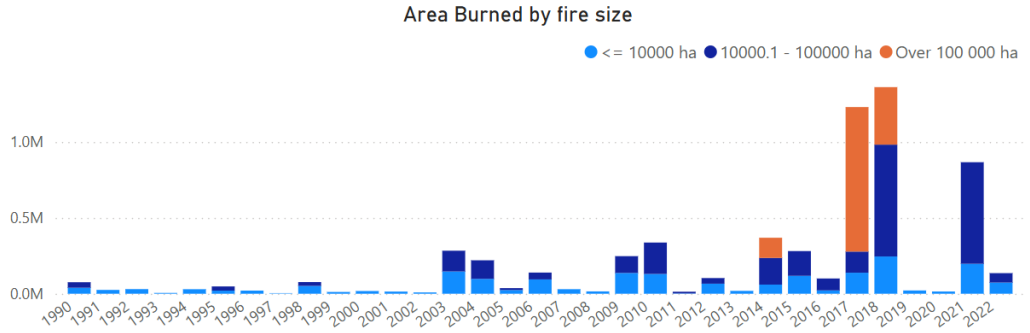
month	Sum of area burned
4	118,779.08
5	319,835.12
6	876,821.75
7	3,485,104.23
8	1,290,704.01
Total	**6,091,244.18**

Next, we examine the annual trend in wildfire extent. When classified by cause, lightning is the leading factor contributing to the largest burned areas. In recent years, there has been a trend of increasing wildfire impact, with notably large burned areas observed in 2017 and 2018. From 1990 to 2022, records show that 56% of all wildfire incidents were caused by lightning.

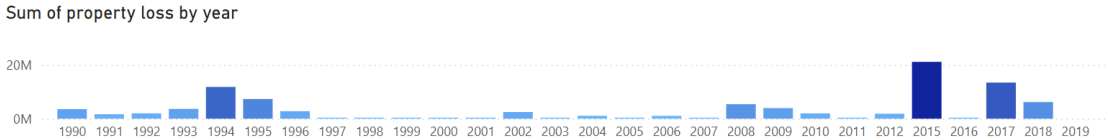




If we categorize wildfires based on the scale of their impact, we can observe that prior to 2014, there were almost no wildfire events exceeding 100,000 hectares in size.



Property loss was highest in 2015 and 2017, reaching approximately 13 million in 2017 and exceeding 20 million in 2015.



Overall, we observe that the impact and scale of wildfires have increased over the past decade. This trend further underscores the importance of predicting wildfire risk and implementing effective prevention strategies.

Modeling

Python approach

Accuracy

With Smote method (on the test data)

Models	AUC	RSME	Accuracy	Precision	Recall
Random Forest	0.9989	0.0853	0.9914	0.9916	0.9780
XGBoost	0.9585	0.1613	0.9659	0.9396	0.9413
Logistic	0.9437	0.2923	0.8763	0.7860	0.7793

Logistic Regression:

- AUC: a 0.9437 score suggests the model distinguishes well between the two classes.
- RSME: on average, the predicted probabilities are off by about 0.2923 from the true class label.
- Accuracy: the model correctly predicted the class for 87.63% of the test data.
- Precision: the model corrects about 79% predicting a positive class (fire occurred)
- Recall: the model correctly identifies about 78 out of every 100 actual positive cases.

XGBoost:

- AUC: 0.9585 score indicates very strong class separation power of the model
- RSME: 0.1613 lower error score than the Logistic model.
- Accuracy: better total prediction that were correct than the Logistic model at 96.59%
- Precision: out of all the instances the model predicted fire occurred, the model corrects about 93.96%

- Recall: the model did way better than the Logistic Regression out of all the positive cases (fire occurred)

Random Forest:

- AUC: outstanding class separation out of the three models used
- RSME: the lowest error score at 0.1613
- Accuracy: the best model overall in predicting power (99.14%)
- Precision: out of all the positive cases (fire occurred) the model has the highest predicting power.
- Recall: again, the model shows consistent predicting accuracy over all metrics (97.80%)

Among the three models evaluated, the Random Forest demonstrates the strongest predictive performance across all metrics. This indicates that it is the most suitable model to apply to the original dataset, especially given the rarity of fire occurrence cases.

Applying the Random Forest to the Original data (without Smote method)

Model	AUC	RSME	Accuracy	Precision	Recall
Random Forest	0.9949	0.0813	0.9942	0.9934	0.9795

- AUC: again, strong class separation power from the model
- RSME: low error at 0.0813 RMSE
- Accuracy: high accuracy level overall at 99.42%
- Precision: the model corrects about 99.34% in predicting the positive cases
- Recall: the model correctly predicted 97.95% of the actual cases.

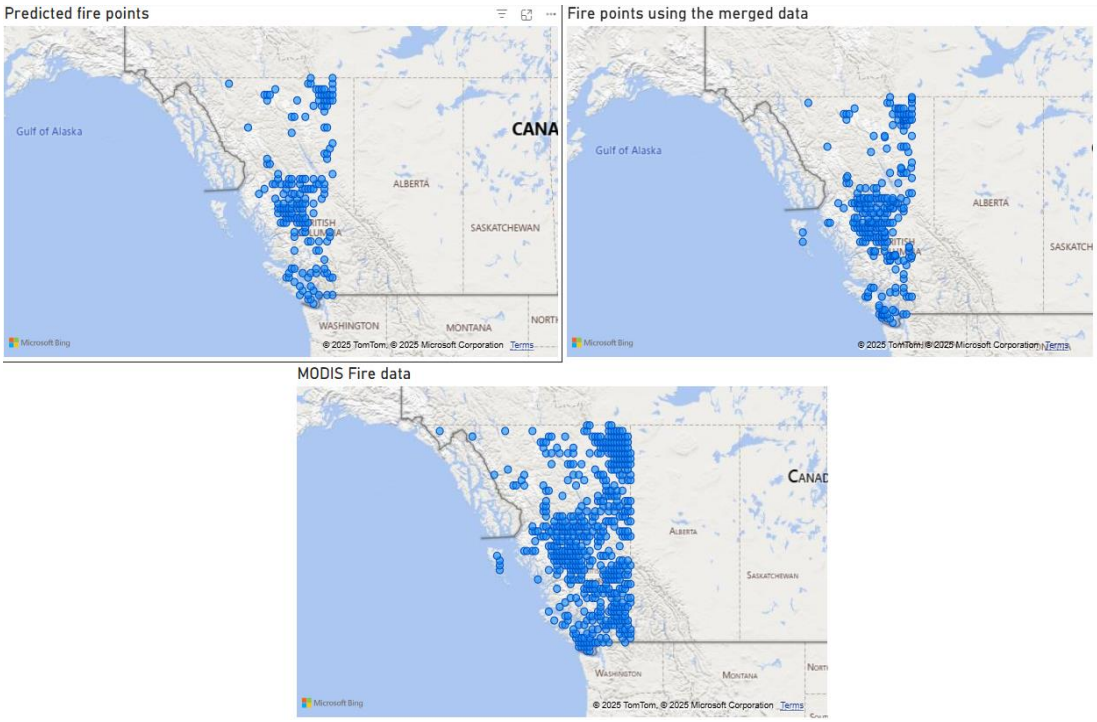
Confusion Matrix (on the original data)

	Predicted Non-Fire Points	Predicted Fire Points
Actual Non-Fire Points	203155	220
Actual Fire Points	1167	33131

The model:

- correctly predicted fire occurrence in 33,131 cases.
- correctly predicted no fire in 203,155 cases.
- wrongly predicted a fire where there was none — a very low false alarm rate.
- The model missed 1,167 cases where a fire occurred.

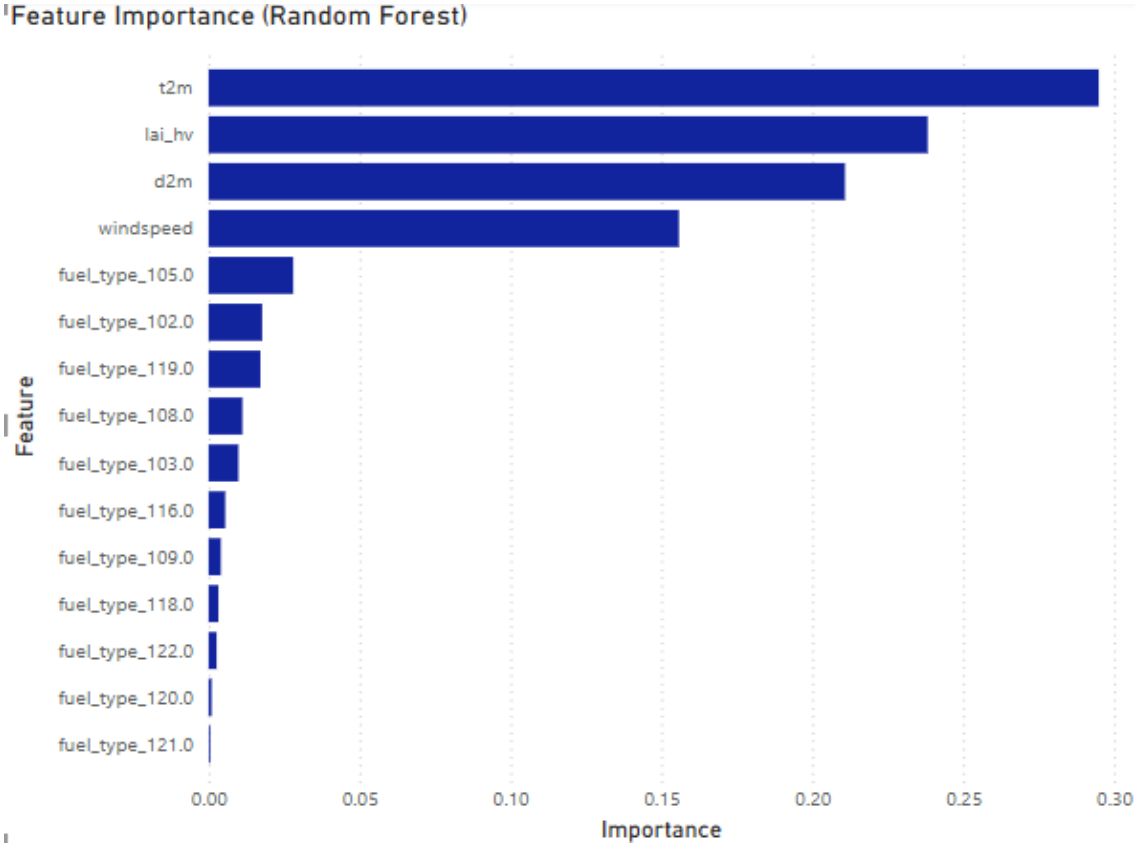
By the visual



The three maps highlight the relationship between the model's predictions and the actual fire data. The top-left map shows the predicted fire occurrences, which are mostly concentrated in central and southern British Columbia, reflecting the model's ability to identify key fire-prone areas. The top-right map

displays the actual fire events from the merged dataset used to train the model; this is the same data source used to generate the predictions in the first map. In contrast, the bottom map presents the complete original MODIS fire dataset, which offers broader spatial coverage, particularly in northern and eastern BC. This comparison illustrates how preprocessing steps—such as rounding coordinates and aggregating to monthly records—while necessary for alignment with climate data, resulted in reduced data granularity and may have limited the model's ability to detect fires in less represented regions.

Feature Importance



The feature importance chart from the Random Forest model indicates that air temperature (t2m) is the most influential factor in predicting fire occurrence. This highlights the critical role that heat plays in fire risk. The second most important variable is leaf area index (lai_hv), suggesting that vegetation density also contributes significantly to fire likelihood — likely due to the amount of available fuel.

Next, dewpoint temperature (d2m) ranks third in importance, which aligns with

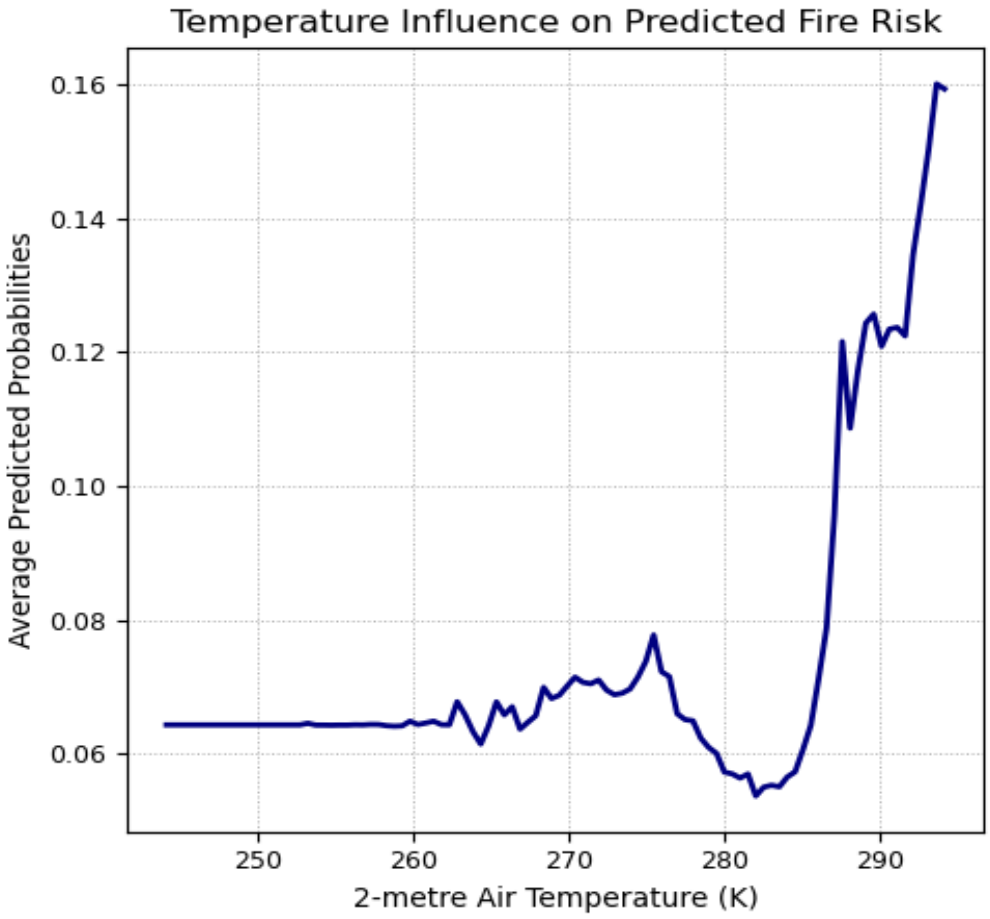
expectations, as moisture levels in the air can influence fire ignition potential. Windspeed is also a notable predictor, albeit slightly less important, reflecting its role in influencing fire spread rather than initiation.

The fuel type features have smaller impacts individually, but they still help the model understand the environment better. Some fuel types, like fuel type 105 (red and white pine) and fuel type 119 (non-fuel), seem to affect the results more than others. This might mean certain types of vegetation or land cover are more likely to be linked with fire and could be looked at more closely in future analysis.

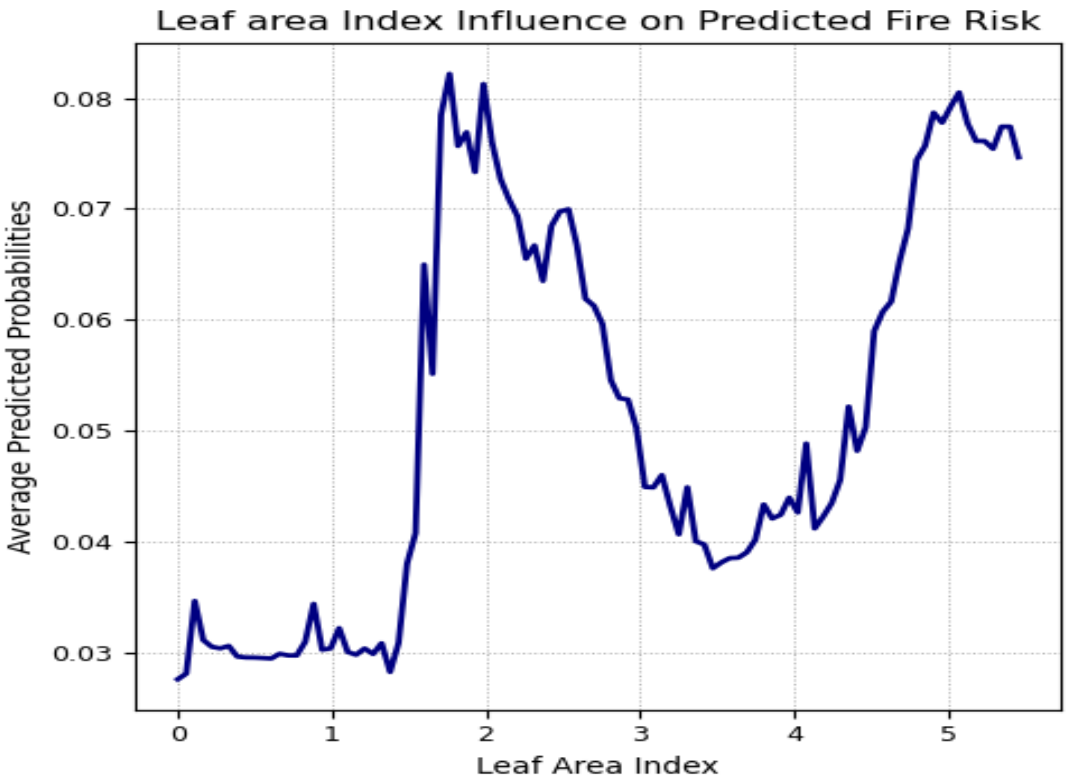
Partial Independence Plot Patterns

To deduce the estimated impact of certain features, partial independence plots were produced which track the average predicted probabilities for each training point with a certain value of the feature. This can be used to isolate the effects of certain variables on a more specific level by normalizing them through the average.

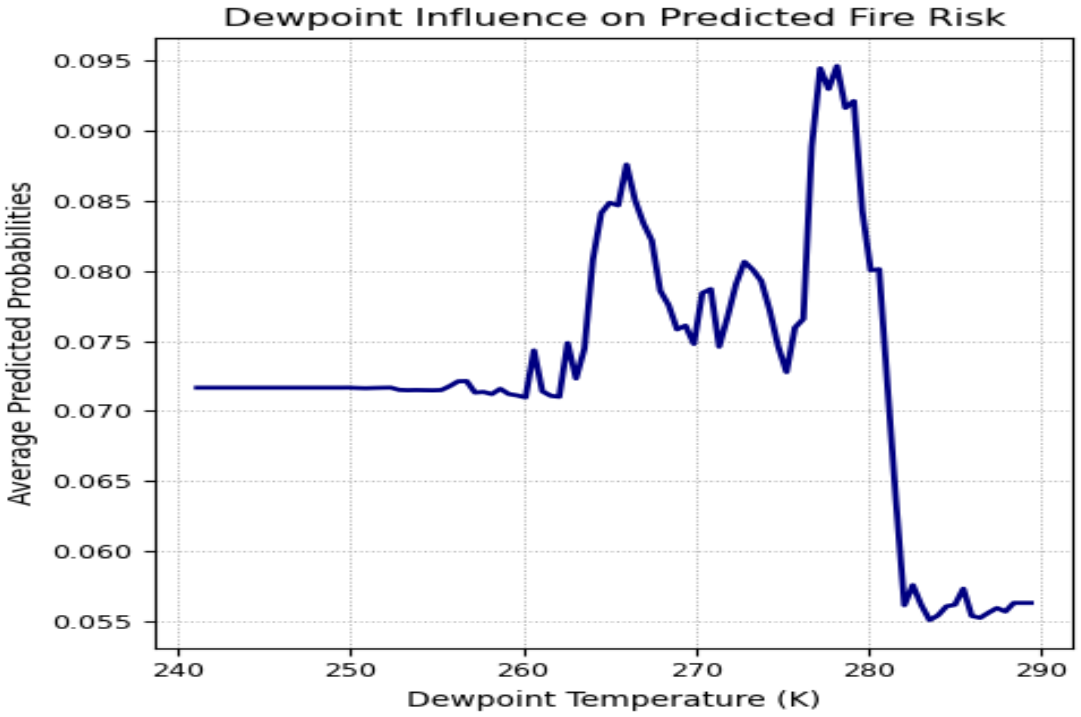
To understand the impact of these variables on our model in more detail, we take the top features ranked by importance to analyze through the variable effect plot:



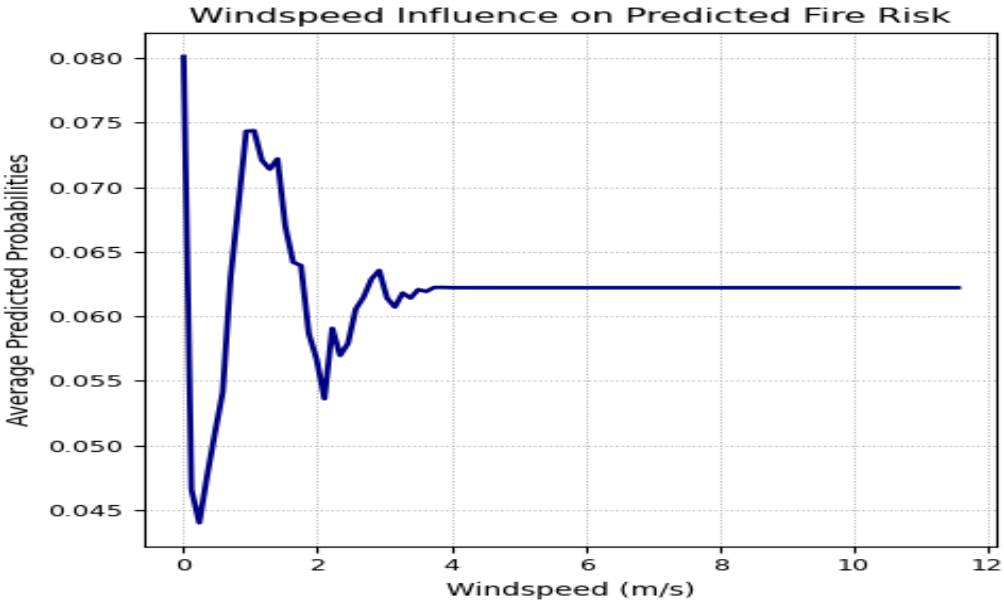
- The chart shows low average probabilities of fire across all t2m values
- However, the probability spikes when the temperature is above the 285 K level.
- This suggests that temperature alone is not sufficient to trigger fire classification, but higher temperatures (above 290K) might cause a significant change to the model.



- Sharp increase in predicted fire probability, peaking around LAI = 2
- Fire probability is quite low and stable over the lens of this variable
- Probability drops from LAI ~ 2- 3.5, suggesting this range may represent denser, moister vegetation (e.g., shaded forests or healthy canopies) that is less prone to fire under typical conditions.



- The predicted fire probability peaks around 275–280 K (roughly 2°C to 7°C dewpoint)
- Moderate dewpoint temperatures correspond to higher fire risk
- The predicted probability drops off significantly after ~ 279 K
- Too dry or too moist air suppresses fire risk. The peak suggests intermediate humidity levels are most associated with fire occurrence



- From about 4 m/s and beyond, the predicted fire probability flattens out

completely indicating that higher wind speeds are not further influencing the model's prediction.

- Thus, in the range above 4 m/s the model does not consider windspeed to have additional predictive power for the fire risk.

Rationale for choosing GIS approach

From the visual map above, we observed a significant inconsistency between the predicted and actual fire occurrence data using the Python approach. This discrepancy highlights a limitation in model accuracy and data alignment. Notably, part of the inconsistency stems from differences in the data formats and resolutions of the two datasets used—namely, the climate data and the fire occurrence records—which may have led to imperfect spatial or temporal merging and contributed to mismatched predictions.

Statistically:

- When we apply the model on the tested data (with Smote)

```
--- Confusion Matrix ---
True Positives (TP): 79564 (27.94%)
True Negatives (TN): 202700 (71.19%)
False Positives (FP): 674 (0.24%)
False Negatives (FN): 1786 (0.63%)
```

- When applying on the original data (without Smote)

```
--- Confusion Matrix ---
True Positives (TP): 33131 (13.94%)
True Negatives (TN): 203155 (85.48%)
False Positives (FP): 220 (0.09%)
False Negatives (FN): 1167 (0.49%)
```

- And with the Modis fire data

```
--- Confusion Matrix ---
 True Positives (TP): 5799 (28.35%)
 True Negatives (TN): 0 (0.00%)
 False Negatives (FN): 14659 (71.65%)
 False Positives (FP): 0 (0.00%)
```

There is a noticeable reduction in the number of both fire and non-fire predictions when the MODIS fire data is compared side by side with the model

output. These statistics confirm that a significant portion of the fire data points did not merge effectively with the climate data, which limited the model's ability to predict fire occurrences accurately when compared to the original MODIS fire data. We recognize this as a major limitation of the current approach and suggest that a GIS-based method may provide a more robust framework for integrating and building the dataset.

```
In [5]: print(monthly_fire.info())
   ...: print(df_climate_filtered.info())
   ...: print("Number of matched records:", len(merged_check))
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 318734 entries, 0 to 318733
Data columns (total 5 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   year           318734 non-null  int32
 1   month          318734 non-null  int32
 2   latitude       318734 non-null  float64
 3   longitude      318734 non-null  float64
 4   fire_occurred  318734 non-null  int64
dtypes: float64(2), int32(2), int64(1)
memory usage: 9.7 MB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1030176 entries, 0 to 1030175
Data columns (total 9 columns):
 #   Column     Non-Null Count    Dtype
---  ------     --------------    -----
 0   u10        1030176 non-null  float32
 1   v10        1030176 non-null  float32
 2   d2m        1030176 non-null  float32
 3   t2m        1030176 non-null  float32
 4   lai_hv     1030176 non-null  float32
 5   year       1030176 non-null  int32
 6   month      1030176 non-null  int32
 7   latitude   1030176 non-null  float64
 8   longitude  1030176 non-null  float64
dtypes: float32(5), float64(2), int32(2)
memory usage: 43.2 MB
None
Number of matched records: 171489
```

Upon reviewing the merged dataset, only 171,489 records remained out of the original 318,734 fire data entries. This further highlights the limitations of the current merging approach and reinforces our decision to adopt a GIS-based method moving forward.

Geospatial approach

Accuracy

We ran 3 models on the data and determined the most accurate one based on these metrics.

Model	AUC	Root Mean Squared Error (RMSE)	Accuracy	Precision	Sensitivity (recall)
Logistic regression	0.8679	0.3479	0.8789	0.9012	0.9596
XGBoost	0.9679	0.2633	0.93066	0.9425	0.9762
Random forest	**0.9883**	**0.1281**	**0.9840**	**0.9890**	**0.9913**

- In all models, the sensitivity was higher than precision, which is indicative of its increased ability to detect fires. This could also be dependent on the sample as 83.1% of the test data consisted of fire points.
- The logistic regression model was the least accurate according to the AUC and RMSE with a value of 0.8679 and 0.3479 respectively. It generally had the lowest values when it came to accuracy, sensitivity, and recall. The AUC was considerably lower than that of the XGBoost and Random Forest model.
- The random forest model was the most accurate out of the 3, having the highest AUC, RMSE, accuracy, precision, and sensitivity.

Confusion matrix

	Predicted Non-Fire Points	**Predicted Fire Points**
Actual Non-Fire Points	122655	7025

Actual Fire Points	5589	633838

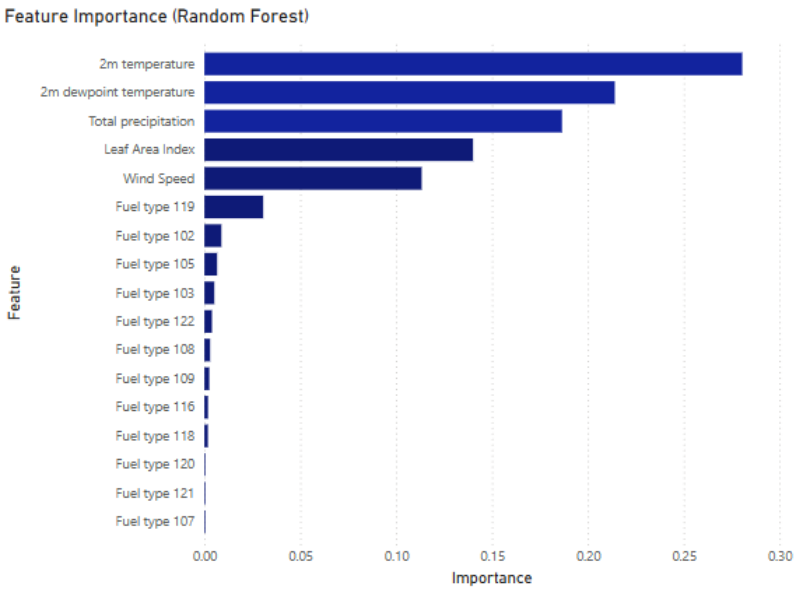
- Out of the 639427 total tested fire points, the random forest model managed to predict 633838 of them accurately.
- Out of the 129680 non-fire points, it managed to predict only 122655 of them accurately.
- The number of fire points predicted inaccurately was less than 1500 more at 7025 than the inaccurately predicted non-fire points at 5589 even though there were almost 5 times as many fire points than non-fire points.
- This reinforces the observation that the model is better at detecting areas with fires than areas without fires.

Map comparison

This outlines the comparison between the predicted fire points and the actual fire points in August 2024 using the random forest model. The map on the left highlights the actual fire points that were predicted correctly, and the map on the right highlights the fire points that were predicted incorrectly.

Out of the 74,819 actual fire points, it managed to predict 74746 of them and failed to predict 73 of them. This highlights the heightened ability of the model to detect fires.

Feature importance

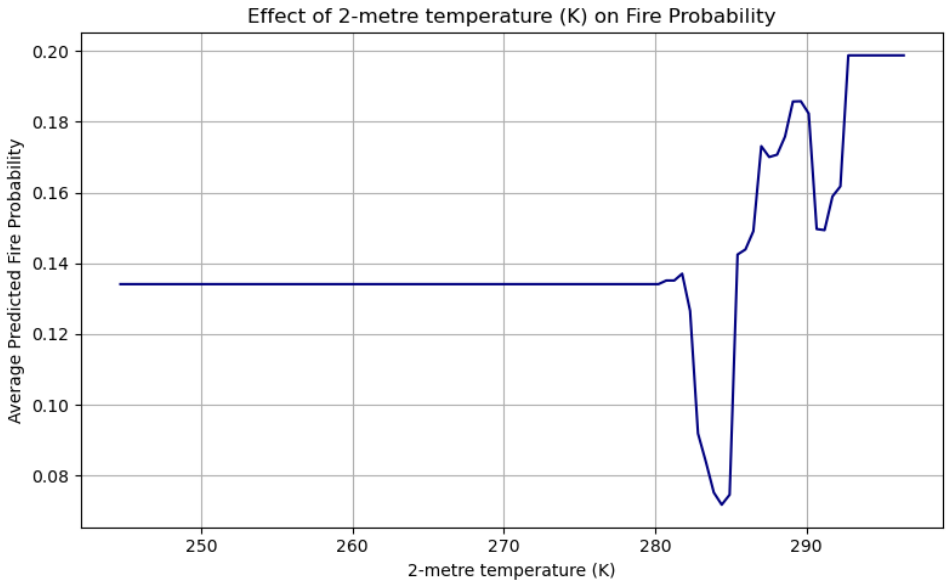


- According to the random forest model, the important features were 2m-temperature (30%), 2-meter dewpoint temperature (21%), and total precipitation (19%). These 3 features account for almost 70 percent of the detection.
- The ordering of the most important features matched the model made by Norton (2024), except for the 2-meter temperature being rated as the 2nd most important based on this rather than the 2nd least important.
- The most important fuel type for the random forest was 119 (non fuel). This type refers to alpine areas with patchy vegetation that would not normally cause fires.

Partial Independence Plot Patterns

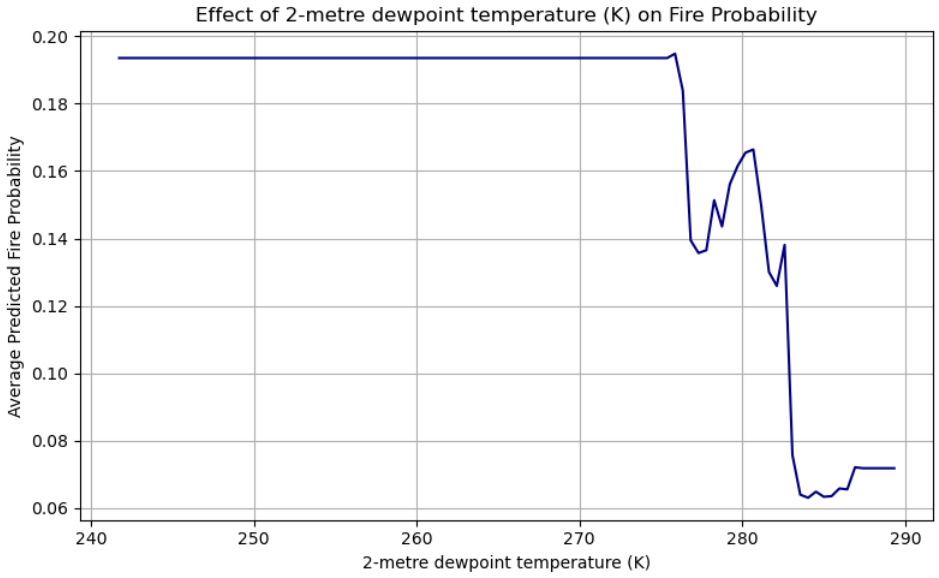
To understand the impact of these variables on the random forest model in more detail, we made partial independence plots based on the top features ranked by importance.

2-meter temperature



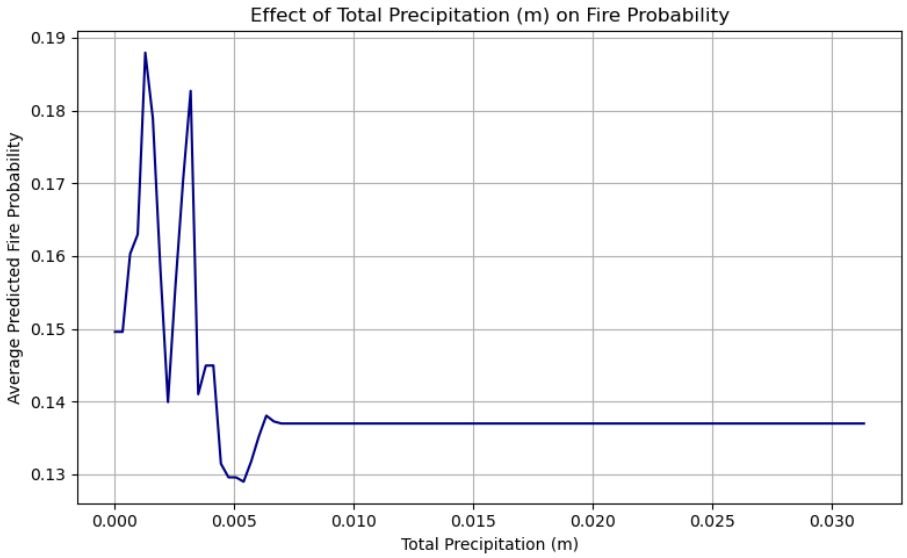
- From 245 K to 282 K, the average probability remains at 13 percent.
- Then it decreases from 13% to 7% from 282 to 284 K.
- From 284 to 285 K, it increases by 7% to 14% and continues to generally rise from there.
- At lower temperatures, there is practically no effect but at higher ones, there is a positive relationship to fire risk.
- This pattern is understandable as high temperatures cause vegetation to dry out, making it more flammable and increasing fire risk.

2-meter dewpoint temperature



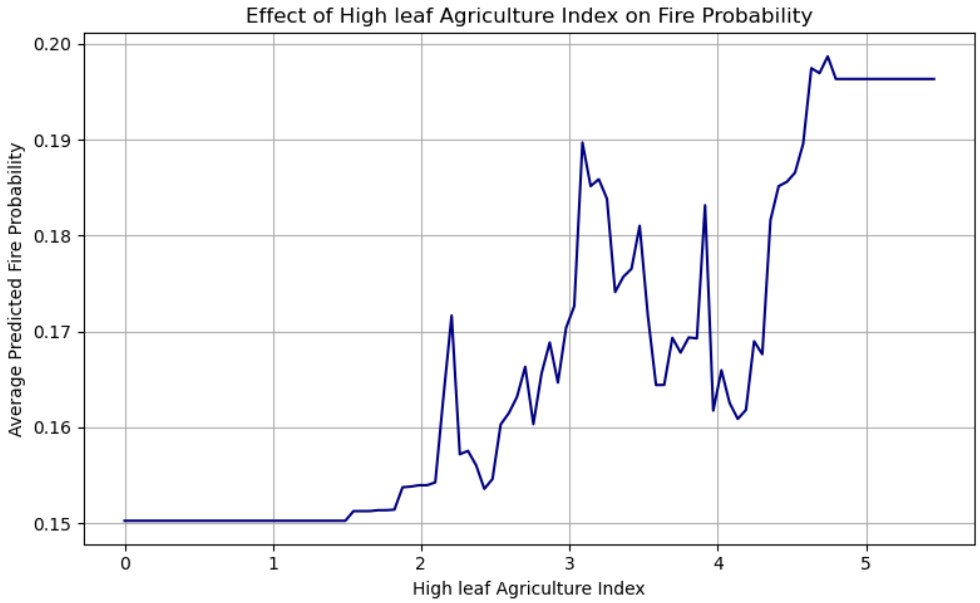
- 2m dew point temperature has an opposite effect on the average probabilities.
- From 241 K to 276 K, it remains at a consistent level of 19%.
- At 277 K, it decreases to 14% and generally exhibits the same trend.
- This is understandable as higher dewpoints indicate more humid air, which decreases the potential for ignition for vegetation.

Total Precipitation



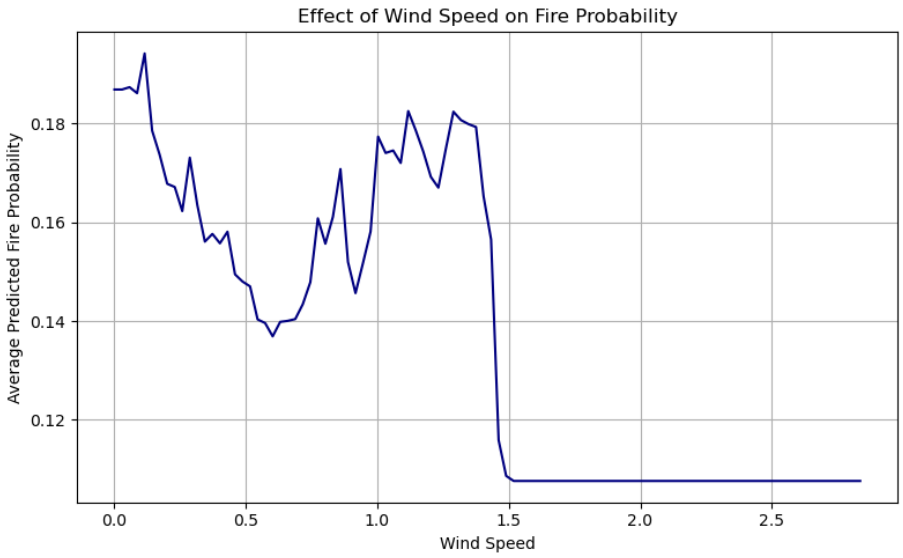
- The average probabilities generally decrease to a level with an increase in total precipitation.
- At extremely low levels ranging from almost 0 to 0.003177 metres, the probabilities fluctuate between 14 and 18 percent generally decreasing in the process.
- At values higher than that, they stay at a consistent level of around 14 percent.

Leaf Area Index



- From 0 to 2.1, the LAI value has a negligible effect on the average area probability
- After that point, it starts to increase until it reaches a peak of 20 percent at 4.63 and before flattening out
- This pattern is understandable with a high LAI indicating more available fuel and more intense fires.

Wind Speed



- The average probability generally decreases until 14 percent at 0.6 m/s before increasing to 18 at 1.32 m/s

- Then there’s a drastic decrease in probability before it levels out at 1.52 m/s. Higher than 1.52 m/s, increasing the wind speed has no effect on the average probability.
- With wind’s potential to spread fire embers over long distances, it would imply that higher windspeeds would be associated with an increase in the average probability, but the average probability was the highest at 0.11 m/s with a value of 20 percent and it was the lowest at 2.83 m/s with a value of 11 percent.

Impact on Property Damage

From the model's feature importance in the 2 approaches, we found that temperature is the most important factor affecting fire risk. To better understand this, we looked at how temperature relates to property damage. This helps explain why we need to focus on managing this variable in implementation plans.

We used data from the National Forestry Database, provided by the client, and combined it with official temperature data from the Government of Canada. After merging the datasets, we created a new database with three key variables: property damage, total area burned (in hectares), and temperature.

```
In [80]: print(impact_df.head(50))
      Year  Total area burned (ha)  Property loss  Mean Temp (°C)
0   1990.0             6315.150000      3569089.0       10.375000
1   1991.0             2099.066667      1675581.0       10.241667
2   1992.0             2537.733333      2001511.0       11.183333
3   1993.0              431.950000      3650159.0       10.066667
4   1994.0             2479.608333     11844924.0       10.783333
5   1995.0             4006.666667      7337770.0       11.116667
6   1996.0             1722.425000      2798557.0        9.683333
7   1997.0              247.316667        17200.0       11.016667
8   1998.0             6396.726167         5028.0       11.308333
9   1999.0              962.950833         1021.0       10.466667
10  2000.0             1484.059333         1088.0       10.250000
11  2001.0             1181.100750          646.0       10.391667
12  2002.0              715.546667      2474250.0       10.483333
13  2003.0            23616.901000        99232.0       11.166667
14  2004.0            18358.038917      1093141.0       11.600000
15  2005.0             3069.414667       102647.0       11.000000
16  2006.0            11618.704333      1071704.0       10.933333
17  2007.0             2498.904500        37012.0       10.333333
18  2008.0             1112.291583      5402869.0       10.245455
19  2009.0            20626.414250      3961725.0       10.408333
20  2010.0            28098.404083      1983140.0       11.166667
21  2011.0             1096.085583        80000.0       10.083333
22  2012.0             8601.564250      1854908.0        8.400000
```

One limitation of this approach is that the final dataset contains only 21 rows. This is mainly because the temperature data we used is only available up to the year 2012, which limits the number of years we could include in the analysis. Additionally, the property damage data is recorded on a yearly basis, while the temperature and area burned data were originally recorded more frequently. To ensure the datasets could be properly merged, we aggregated the temperature and area burned data by calculating their yearly averages. This allowed us to align all three variables—property damage, area burned, and temperature—on a yearly level for consistent and accurate comparison.

	Total area burned (ha)	Property loss	Mean Temp (°C)
Total area burned (ha)	1.000000	-0.073957	0.319637
Property loss	-0.073957	1.000000	-0.017916
Mean Temp (°C)	0.319637	-0.017916	1.000000

After merging the datasets, we examined the relationships between total area burned, property loss, and mean temperature using a correlation matrix. As shown in the table, the correlations between these variables are relatively weak. For example, the correlation between mean temperature and total area burned is about 0.32, indicating a mild positive relationship. Meanwhile, the correlation between property loss and the other two variables is close to zero and slightly

negative, suggesting no meaningful linear relationship. These results show that, at least in this dataset, there is no strong or direct correlation between temperature, area burned, and property damage. This highlights the complexity of wildfire impacts and suggests that other factors may also play a significant role in driving property loss beyond just temperature or burned area.

We then moved on to the second approach, where we merged the temperature and property damage data with the final dataset used for our prediction models. This allowed us to explore a wider range of variables and gain a broader view of the factors influencing fire risk. Using this expanded dataset, we applied the same correlation matrix method to study the relationships between temperature, property damage, and the other variables included in the model.

```
In [93]: print(corr_matrix)
                    d2m       t2m    lai_hv  windspeed  Property loss
d2m            1.000000  0.885909  0.499743  -0.424907       0.364856
t2m            0.885909  1.000000  0.653315  -0.549162       0.499363
lai_hv         0.499743  0.653315  1.000000  -0.889463       0.451172
windspeed     -0.424907 -0.549162 -0.889463   1.000000      -0.326728
Property loss  0.364856  0.499363  0.451172  -0.326728       1.000000
```

From the updated correlation matrix, we can see a clearer picture of how different variables relate to property damage. Among all the features, **t2m** (mean temperature at 2 meters) shows the highest positive correlation with property loss, with a value of **0.50**. This further supports our earlier finding that temperature plays a key role in fire impact. Other variables, such as **d2m** (dew point temperature) and **lai_hv** (leaf area index high vegetation), also show moderate positive correlations with property damage, at **0.36** and **0.45** respectively. On the other hand, **windspeed** has a negative correlation of **-0.33**, suggesting that higher wind speeds may be associated with lower property damage in this dataset. Overall, this analysis highlights that temperature remains the most influential factor when it comes to predicting property loss from wildfires.

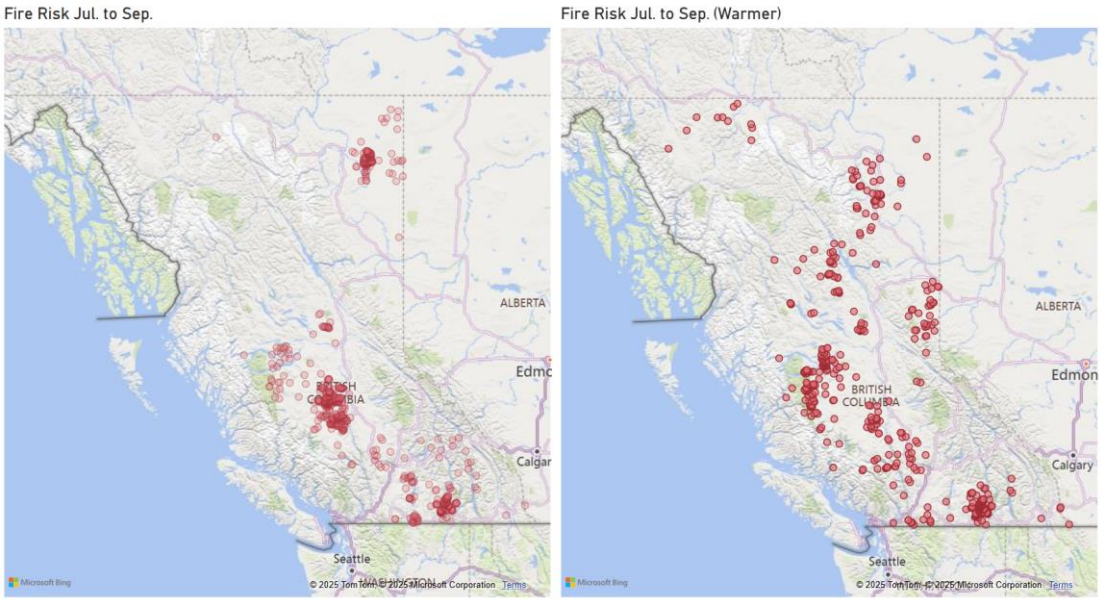
Seasonal risk Prediction

We now turn to seasonal risk prediction, aiming to assess wildfire likelihood and

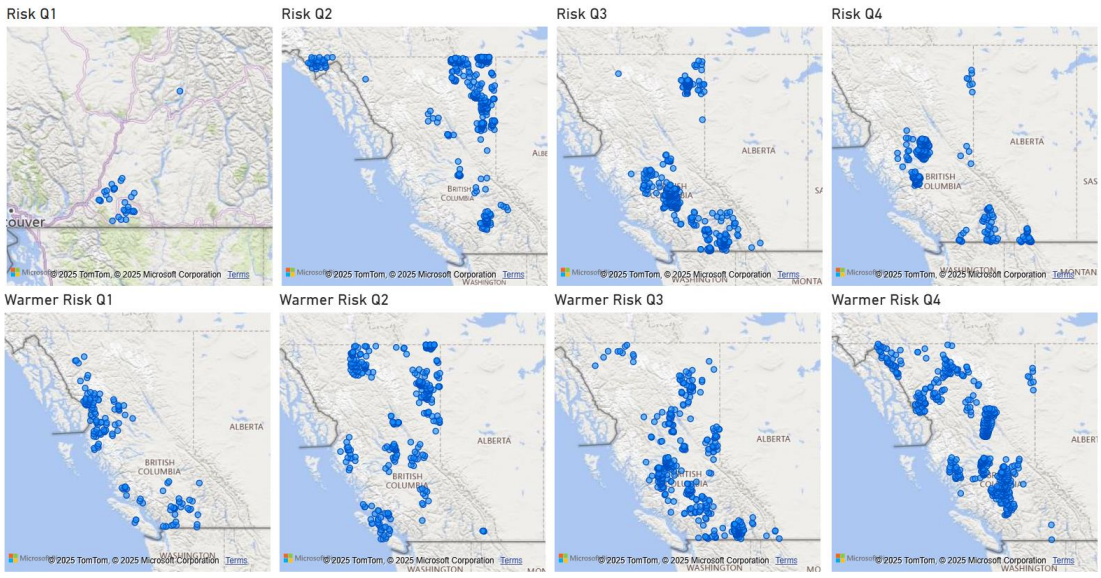
potential severity across different times of the year.

To estimate the seasonal probability of wildfire occurrence in 2025, data from 2000 to 2024 were used to simulate the 2025 scenario. British Columbia was divided into multiple sub-regions to calculate seasonal climate information for each area. As a baseline, simulations were first conducted using the historical average values over the reference period. The prediction model employed is a GIS-based Random Forest model. Since 2m temperature and 2m dewpoint temperature are identified as the two most important features in the model, an alternative scenario—referred to as the "warmer" scenario—was created by adjusting these two variables.

The following two figures present the risk prediction results for the third quarter. The left figure shows the predicted risk under average seasonal climate conditions, while the right figure displays the results under the warmer scenario, in which the temperature features were increased by one standard deviation. It can be observed that under the warmer scenario.



After multiple simulations, we found no significant difference in the total number of predicted wildfires between the two scenarios. However, in the warmer scenario, the predicted fire occurrences are more geographically dispersed. The figure below presents the predicted wildfire risk across all four quarters of 2025.



Key Findings

Based on our analysis, the 2-metre temperature, 2-metre dewpoint temperature, and total precipitation were the most important features in order for the geospatial model. In contrast, for the Python-based model, the leaf area index was the 2nd most important feature with the 2-meter dewpoint temperature being 3rd. The total precipitation was not seen as important to any degree in the Python-based model.

Based on the average probability graphs, 2-meter temperature increases generally correspond to increases in fire probability for both models. The opposite relationship is seen for the 2-meter dewpoint temperature. In the geospatial model, total precipitation increases correspond to decreases in the fire probability. In the Python model, there are probability decreases at higher leaf area index values. The wind speed does not seem to have any noticeable pattern with varied average probabilities based on the initial speeds.

The potential impact of temperature is highlighted in the correlation matrix between the climate variables and property damage. Both the 2-metre and the dewpoint temperature variables have a positive correlation with the high vegetation leaf area index also having a positive correlation with property damage.

Based on the seasonal risk prediction, increasing the temperature variables leads to an increase in the number of fire occurrences. Though after multiple simulations, there was no significant difference in the total number of predicted wildfires between the two scenarios. Notably, in the warmer scenario, the predicted fire occurrences were more dispersed geographically.

Recommendations

Based on the analysis, it is important to monitor 2-meter temperature values daily with a drone to capture places that are likely to hit the 280 K mark for the 2-metre temperature as the highest average probability jumps occur in that range based on the geospatial Random Forest model.

Based on the random forest analysis, monitor 2-metre dewpoint temperatures up to 280 K as average fire probabilities are higher. It's important to monitor places with low precipitation totals as on average, they have a higher chance of fires.

In addition, based on the correlation matrix, it's also important to capture areas with a high vegetation leaf area index as they are positively correlated with high property damage values.

With the appropriate information, BC SMART can take certain actions proactively to account for the effects of the fires in areas where the forecasted risk is high.

Costs

Due to the large number of data points collected, one would need to purchase a device with a high processing power to run the code accurately whilst being efficient in the run time. One would need specific high-end drones which specialise in climate surveillance. These can cost up to $22,000 to buy and use depending on the area covered in surveillance. Time would also need to be invested in collecting data. Finally, there will be implementation costs that would be associated with testing out the different strategies in search for the right

response based on the risk.

Benefits

Despite the costs, by looking out for certain details, and monitoring the risk, wildfire services and local utilities are able to take certain measures proactively to limit the impact of wildfires on infrastructure and the environment.

Implementation plan

BCIT SMART can use these models to create maps highlighting the predicted probabilities. With the existing climate data, they can create 5-year time series forecasts on each of the climate variables across BC and use the random forest model to generate predicted fire probability maps based on the forecasted climate variables.

A year prior, they can send drones to survey power utility infrastructures with a predicted probability of 50 percent and above to track the temperature and precipitation features of those high-risk areas in more detail to make forecasts and fire predictions one season ahead.

Based on the risk level, wildfire services and local utilities will adjust their operations and implement different strategies such as crew safety bulletins or direct public safety power shutoffs.

Conclusion

With the right model, one can accurately predict the occurrence of fires. This can be further implemented with necessary forecasting on the relevant explanatory variables to detect areas with a high risk of fires one season ahead. Although it is costly to implement due to the resources required, it will ultimately save lives and reduce the impact wildfires have on the environment and property.

References

1. Canada's National Forestry Database: http://nfdp.ccfm.org/en/data/fires.php
2. **Environment and Climate Change Canada.** *Canadian Climate Data – Monthly Data Report for Station ID 1309 (Vancouver, BC).* Government of Canada, 2024. Available at: https://climate.weather.gc.ca/climate_data/monthly_data_e.html
3. **Fetz, T., Jones, I. L., Wilhelm, S. I., Kouwenberg, A.-L., & Ramey, A. M.** *Comparing cost-effectiveness of radio and drone telemetry with playback surveys for assessing translocation outcomes.* Vol. 62. *Journal of Applied Ecology*, 2025.
4. **Hersbach, H., Bell, B., Berrisford, P., et al.** *ERA5 monthly averaged data on single levels from 1979 to present.* Copernicus Climate Change Service (C3S), Climate Data Store (CDS), 2023. Available at: https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels-monthly-means?tab=download
5. McNorton, J. R., et al. "A Global Probability-Of-Fire (PoF) Forecast." *Publication Title*, 14 June 2024.
6. Mohajane, Meriame, et al. *Application of Remote Sensing and Machine Learning Algorithms for Forest Fire Mapping in a Mediterranean Area*. Ecological Indicators, vol. 129, Oct. 2021.
7. **NASA FIRMS.** *MODIS Active Fire Data.* NASA Earth Science Data and Information System (ESDIS), Fire Information for Resource Management System (FIRMS). NASA, 2024. Available at: https://firms.modaps.eosdis.nasa.gov/download/create.php.
8. Van Wagner, Charles E. *Structure of the Canadian forest fire weather index*. Vol. 1333. Ottawa, ON, Canada: Environment Canada, Forestry Service, 1974.

Appendices

Data Sources

We used Canada's National Forestry Database to analyze historical trends on burned areas and fire occurrences. It is relevant to provide an appropriate overview of the situation in BC. More information can be found in the data sources section of the appendix.

We also obtained fire data from National Resources Canada from 2000 to 2024 in the form of hotspots. Each hotspot data file contains information on where and when each fire occurred. It provided a starting point to understanding fire characteristics. More information can be found in the Data Sources section of the appendix.

In the preprocessing step, we used ERA5 monthly data, which only provides values at the 00:00 UTC time frame. As a result, we must assume that variables such as dew point temperature and air temperature are representative of the broader daily and regional conditions. However, users should be aware that this introduces limitations and are encouraged to explore alternative or higher-resolution datasets to improve accuracy and granularity of the analysis.

Methodology

Geospatial Approach

To ensure appropriate sampling, the number of randomly generated points was made to be equal to the number of fire hotspots that month. On months without fires, they were based on the monthly average number of fire hotspots in that year.

One limitation was that in earlier years, on months with fires, the number of randomly generated points was equal to the number of fire hotspots, but in later years, QGIS was unable to generate the appropriate number of non-fire points. This ended up skewing the data to have more fire hotspots than non-fire

generated points. This was due to a limit placed in the code to ensure adequate processing time. This limit can be avoided using a system with greater processing power.

Python Approach

There are a few limitations for the data preparation:

```
In [12]: df_fire.head()
Out[12]:
   latitude  longitude  brightness  scan  track   acq_date  acq_time  \
0   49.2088  -118.7312       304.0   1.1    1.0 2000-11-01       611
1   49.6365  -119.3867       309.0   1.1    1.0 2000-11-01       611
2   49.2070  -118.7463       306.3   1.1    1.0 2000-11-01       611
3   49.4978  -120.6016       301.9   1.0    1.0 2000-11-01       611
4   49.4886  -120.6698       323.4   1.0    1.0 2000-11-01       611

  satellite instrument  confidence  version  bright_t31   frp daynight  type  \
0     Terra      MODIS          58     6.03       271.1  13.1        N     0
1     Terra      MODIS          76     6.03       266.6  16.5        N     0
2     Terra      MODIS          67     6.03       271.2  14.5        N     0
3     Terra      MODIS          45     6.03       270.7  10.3        N     0
4     Terra      MODIS         100     6.03       273.4  27.1        N     0

   year  month  fire_occurred
0  2000     11              1
1  2000     11              1
2  2000     11              1
3  2000     11              1
4  2000     11              1
```

First, the MODIS fire data includes coordinates in float format with four decimal places, while the climate data is organized on a grid with 0.25-degree increments for both latitude and longitude. To address this mismatch, we rounded the fire data coordinates to the nearest 0.25 degrees to align with the climate data grid.

Additionally, we created a new column called **"Fire_occurred"** in the MODIS dataset and assigned a value of **1** to each observation, indicating that a fire event took place at that location and time. This labeling allowed us to later merge the fire occurrence data with the climate variables and use it as the target variable for classification modeling.

```
# Convert longitude from 0–360 to –180–180 if needed, then snap
df_climate['longitude'] = df_climate['longitude'].apply(lambda x: x if x <= 180 else x - 360)
```

Another important limitation we encountered involved the format of the

longitude values in climate data. The ERA5 climate dataset records longitudes in the range of 0 to 360 degrees, which is a common format for meteorological data. However, most geographic datasets—including our fire data—use the standard format of -180 to 180 degrees. To ensure both datasets could be accurately merged based on geographic location, we applied a transformation to convert the longitudes in climate data.

```
In [13]: print(df_climate.head())
  valid_time  latitude  longitude  number expver    lai_hv       u10  \
0 2000-01-01      60.0    -140.00       0   0001  1.218750 -1.050616
1 2000-01-01      60.0    -139.75       0   0001  2.361328 -0.938800
2 2000-01-01      60.0    -139.50       0   0001  1.373169 -0.909503
3 2000-01-01      60.0    -139.25       0   0001  0.229004 -0.885089
4 2000-01-01      60.0    -139.00       0   0001  0.001953 -0.572101

        v10         t2m         d2m        tp  year  month
0 -1.860607  265.404419  259.015930  0.009583  2000      1
1 -1.433849  264.002075  257.586243  0.009347  2000      1
2 -1.252697  262.482544  256.685852  0.008894  2000      1
3 -1.089611  260.949341  255.816727  0.008427  2000      1
4 -0.984631  259.904419  255.183914  0.008264  2000      1
```

Third, merging the two datasets required rounding the fire data coordinates to match the grid-based format of the climate data. We understood that this step would result in a loss of spatial precision and potentially reduce the number of usable data points. Despite this limitation, we chose to proceed to test how well the merged dataset would support our modeling efforts and to evaluate whether meaningful insights could still be drawn from the combined data.

```
In [15]: print(merged['fire_occurred'].value_counts())
fire_occurred
0    1016872
1     171489
Name: count, dtype: int64
```

The final limitation we observed after merging the datasets was a significant imbalance between fire and non-fire observations. As shown in the figure, there are 1,016,872 instances where no fire occurred (fire_occurred = 0) compared to only 171,489 fire events (fire_occurred = 1). This large gap can cause prediction models to become biased toward the majority class, meaning they are more likely to predict "no fire" simply because those cases dominate the dataset.

To address this issue, we applied the SMOTE (Synthetic Minority Over-sampling

Technique) method. SMOTE works by generating synthetic samples for the minority class—in this case, fire occurrences—based on existing data patterns. This helps to balance the dataset and give the model a more equal representation of both classes during training.

```
In [24]: print(balanced_smote['fire_occurred'].value_counts())
fire_occurred
0    1016872
1     406748
Name: count, dtype: int64
```

Now, the non-fire data accounts for about 40% of the whole dataset. It's important to note that we only used the SMOTE-balanced data for training and evaluating our prediction models, in order to reduce bias and improve the model's ability to detect fire events. Once we identified the best-performing model, we then applied it back to the original (imbalanced) dataset for final predictions. This approach ensures that our model is both fair and realistic when used in real-world conditions where fire events are naturally rare.