Wildfire Prediction and Analysis

Serena Dhillon

CMPT 353 Summer 2025, Prof. Greg Baker

Accomplishment Statement Conducted data collection, cleaning and storage to support wildfire analysis using tools like Spark, HDFS and Pandas for weather and fire data. Performed analysis using correlation analysis, chi-squared test, and machine learning techniques to investigate four research questions. Identified a significant relationship between temperature, precipitation and fire count. Trained a regression model explaining 32% of fire occurrences only using weather.

1 Introduction

Canada is known for its breathtaking scenery, from the beautiful landscapes, to the lakes and forests that this country has from coast to coast. Yet every year across Canada, on average 2.5 million hectares of our great forests burn [1]. This impacts both our wildlife and various communities, from drastically declining air quality to entire communities being forced to evacuate due to nearby wildfires.

Our firefighters and equipment have been running dry in an attempt to save our forests and our homes. As of August 2025, there are currently 711 wildfires burning throughout Canada, adding to the total of 4,325 fires this year alone and 7.3 million hectares burned [2, 3]. As the years go by, temperatures continue to rise, and the frequency of wildfires has increased. This begs the question, what relation does our weather have when it comes to wildfires, and it is possible to use this to predict the severity or location of these fires? With records of where firefighters are being sent out, where equipment is going, and the cities and towns needing to be evacuated, what if we could mitigate this risk by predicting risk by area?

This report explores the use of data science to combine wildfire and weather analysis in the hopes to provide early information to fire teams throughout the country in an effort to help improve response times and wildfire management.

For this report we investigated the following questions:

1. How does the average maximum temperature per month and the monthly total of precipitation indicate fire size?
2. How does the average maximum temperature per month and the monthly total of precipitation affect the amount of fires?
3. Based on the location, can we predict the number of fires per month in that area?
4. Based on the location, can we predict the hectares burned within that area?

2 Materials

To conduct our analysis and answer the research questions, we gathered data from multiple online sources, focusing on historical wildfire occurrences and corresponding weather data in Canada. The data sources include government datasets, open data repositories, and academic resources.

Weather Data The historic weather data used for this report was extracted with the help of SFU's Compute Clusters as there mere volume of data these compute clusters was too much to be able to do any other way [4]. We used big data tools such as Spark and the Hadoop Distributed File System to extract, clean and merge this data before loading in onto a single computer to do processing and analysis [5].

National Forestry Database The National Forestry Database held information about the amount of fires per year and the amount of fires that occurred based on different causes [6]. All this data was separate by province and territory and downloaded cleanly into a CSV for processing.

Canadian Wildfire Information System The Canadian Wildfire Information System had statistical information of how many hectares burned per month and situation reports together for information gathering [7]. The data from this source did download into a CSV but was not used for analysis in this report.

Wildfire Dataset (Kaggle) As the main goal of this report was to be able to give analysis by area, location data was required [8]. The data included in this wildfire dataset provided necessary location data to complete analysis. The Government of Canada and Province websites unfortunately did not have for public access.

3 Methods

As our data came from multiple sources, a range of techniques were used to clean and conduct analysis in hopes to answer our questions.

3.1 Extract, Transform and Load

Majority of the data used in this project was obtained by downloading CSV files from the sources listed above. However, extracting the weather data required a more involved process, due to the large volume of data. Using HDFS and Spark we were able to extract data from the necessary weather stations in Canada, and take only the pieces of information required (which in this case was maximum temperature and precipitation). We also grouped the data by month, taking the average and sum respectively, as it would be easier to manage the fire data. Aggregating to the monthly level helped reduce the impact of missing days, minimized daily noise, and allowed us to better observe long-term trends and patterns in the data.

Once we pulled all our data into pandas, we prepared it for analysis. This process included changing the names of the Provinces and Territories to be consistent throughout the datasets and to include Provincial Parks to be within those categories as well, since not all datasets differentiate them. We also grouped the data by month and year, so that time was consistent across datasets. Additionally, we dropped data from before 1990 to save on processing time.

After completing data cleaning for all four files, it was soon realized that merging the datasets without the latitude and longitude locations to the weather data which did not have Province names was not possible. We decided to stick with the weather data and the Wildfire Dataset. These datasets had both the latitude and longitude locations in which we calculated the distance of each fire to all the locations of the weather stations to match the best weather information to the fire, successfully merging our datasets.

3.2 Analysis

Correlation Analysis A correlation analysis was to look at the relationships between maximum temperature, precipitation and fire size to answer our first question. We used both a Spearman and Pearson correlation to do so to show the difference between the nonlinear and linear relationship between these features.

Chi Squared To address our second question, we used a Chi-Squared analysis. Since this is a more categorical approach we first split our variables in question (average maximum temperature and total precipitation) into 3 categories. A chi-squared test was performed on the contingency table created from these categories to determine if there's a significant relationship between temperature and precipitation in the context of fire count.

Regression For the third and fourth questions, we performed Random Forest Regression and K-Nearest Neighbors Regression to predict the number of fires and the total hectares burned within a given area. Input features included spatial, temporal, and environmental variables such as latitude, longitude, temperature, precipitation, and time of year.

Categorization Based on our results with the regression we decided to redo the last question in terms of categories, splitting our hectare size into 3 categories and attempting to predict which category they are in based on the different spatial, temporal, and environmental features. For categorization, Random Forest, K-Nearest Neighbor and Naive Bayes Classifier were used.

4 Results

4.1 Weather Indications of Fire Size

To analyze how weather affects fire sizes, we first decided to take a look at a scatter plot of the two weather features we chose to analyze, the average maximum temperature and the total precipitation per month. We can see this below in Figures 1 and 2.

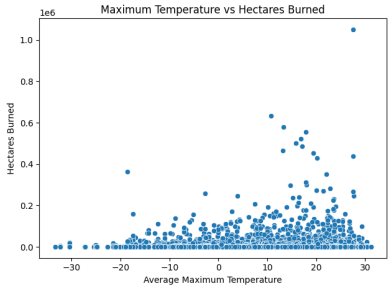

Fig. 1. Temperature vs Hectares

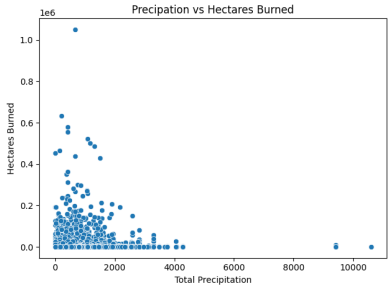

Fig. 2. Precipitation vs Hectares

Although most data points cluster near the bottom of each graph, some patterns are still observable. In Figure 1 we can see how there is more clustered as the temperature increases, and more hectares burned. While in Figure 2 we can see more hectares being burned as precipitation decreases. Since the scatter plots suggest some patterns, to further investigate, we analyzed the Pearson Correlation and Spearman Correlation. This can be seen in Figures 3 and 4. As we can see in the graphs when looking at the relationship of maximum temperature and precipitation to the number of hectares burned, there is a very small linear relationship according to Figure 3 and for the Spearman Correlation there is a very small positive correlation between temperature and hectare size. Given how small the correlation values are, we conclude that weather variables alone are not strong predictors of fire size in this dataset.

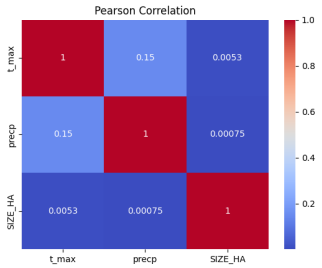

Fig. 3. Pearson Correlation

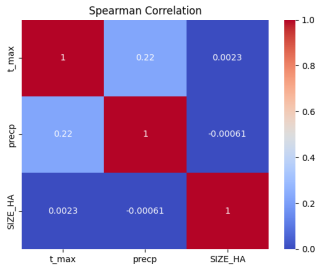

Fig. 4. Spearman Correlation

4.2 Weather Indications of Number of Fires

When analyzing how weather indicates the number of fires we took a slightly different approach. We conducted a Chi-Squared Analysis for this section. We first counted the amount of fires that occurred per month across Canada and created bins for the temperature and precipitation (high, medium and low). The resulting chi test had a p-value of 3.25×10^{-16}. This indicates that this is a statistically significant relationship between the two weather features and the number of fires. Hence the number of fires that occur throughout Canada is not independent of temperature and precipitation levels.

4.3 Predicting Number of Fires

Now that we know that weather indicates the number of fires, are we able to predict how many fires will occur for a given area based on the weather conditions. For this we decided to use a Random Forest Regressor which is a good method to predict for non linear relationships. We trained a model using 450 estimators and a maximum depth of 30, which achieved a score of 0.32 on the validation set, indicates about 32% of fires could be explained with the weather features used. This result is expected considering we are only using two weather features for this analysis and we were not considering non-natural causes of wildfires in this analysis.

4.4 Predicting Fire Size

For this section, although the correlation analysis done earlier does not give a strong indication that weather is correlated to fire size, we decided to still look at the possibility of prediction given the small positive result. We looked at both a regression result and a classification result in attempts to make it slightly easier to predict.

Regression In terms of regression, we used both a Random Forest Regression and a K-Nearest Neighbours Regression. Knowing that there was not a strong correlation between the weather features focused on and fire sizes, we weren't expecting great results. This followed in our outputs with both regression tools giving us negative numbers, -0.15 and -0.05 respectively, suggesting that they underperformed.

Categorization In an attempt to make prediction easier we decided to predict whether the fire was either 'small', 'medium' or 'large' in terms of hectares burned. Given the very little positive correlation seen between our weather conditions and fire size, these results turned out better than expected. As seen in Table 1 below, all the models performed quite well, having accuracy above 0.40 in all three models. This suggestions that our models were not just randomly guessing, and can perform quite well.

Class	Metric	Random Forest	KNN	Bayes
	Accuracy	0.45	0.42	0.42
Large	Precision	0.53	0.47	0.42
	Recall	0.55	0.54	0.48
	F1-score	0.54	0.50	0.45
Medium	Precision	0.36	0.35	0.37
	Recall	0.34	0.34	0.31
	F1-score	0.35	0.35	0.34
Small	Precision	0.47	0.45	0.46
	Recall	0.47	0.39	0.47
	F1-score	0.47	0.42	0.46

Table 1. Model Performance of Fire Size Predictions

While the scores were not very high, they suggest that weather features contribute useful information for predicting fire size, though they are not sufficient to make accurate predictions by themselves.

5 Limitations

Although we had some successes, there were notable limitations that occurred along the way. One major challenge was the limited data available. We were hoping for more weather data such as average wind speed, evaporation, sun exposure, etc., but when extracting information from the cluster there was not enough data on those features to be able to accurately predict with. Another challenge included our data limitations, not all the data collected had either or both of jurisdiction or coordinate data which prevented us from successfully merging the data in a clean manner.

6 Conclusion

In this paper, we aimed to answer four questions: How does average maximum temperature and total precipitation indicate fire size? Do those features affect the amount of fires? Are we able to accurately predict the number of fires per month by area and weather conditions? And are we able to predict the hectares burned within an area based on weather? We used data from the GHCN weather clusters, a Kaggle dataset and some government sites to clean and analyze in hope to answer these questions.

We were able to conclude a statistically significant relationship between our two weather features and the number of fires within a given area (as p ¡ 0.05), and were able to predict 32% of fires with these weather features, which was expected since we did not consider all wildfire causes. Although in the end, we could not conclude that those weather features were an indication of the size of a fire. We were surprisingly able to categorize the fires into a 'small', 'medium', or 'large' class, which indicates that these may be features that could be used in combination with other features to successfully predict fire size.

7 Future Work

Our limitations present valuable opportunities for future improvements in both this project and the broader field of wildfire analysis. Additional data could further help this research develop further, such as adding population data to help classify possible types of fires, adding forest density to look at burn rate to help predict hectares burned, etc. Pulling in other types of data can help with the overall prediction of amount and size of fires by location, and would be beneficial to help fire teams across the country better understand the possibility of what they are dealing with to help them strategize for what is predicted to come.

References

[1] Canadian Red Cross. *Wildfires: Information & Facts*. https://www.redcross.ca/how-we-help/emergencies-and-disasters-in-canada/types-of-emergencies/wildfires/wildfires-information-facts. Accessed: Aug 11, 2025. 2025.

[2] Canadian Interagency Forest Fire Centre. *Canadian Interagency Forest Fire Centre*. https://www.ciffc.ca/. Accessed: Aug 11, 2025. 2021.

[3] Government of Canada. *Natural Resources Canada*. https://cwfis.cfs.nrcan.gc.ca/report/graphs. Accessed: Aug 11, 2025. 2021.

[4] NOAA. *Global Historical Climatology Network*. https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily. Accessed: Aug 11, 2025. 2021.

[5] Greg Baker. *GHCN Dataset*. URL: https://github.sfu.ca/ggbaker/cluster-datasets/tree/main/weather.

[6] *Forest Fires*. URL: http://nfdp.ccfm.org/en/data/fires.php.

[7] *Canadian National Fire Database*. URL: https://cwfis.cfs.nrcan.gc.ca/ha/nfdb.

[8] Ulas Ozdemir. *Wildfires in Canada (1950-2021)*. 2021. URL: https://www.kaggle.com/datasets/ulasozdemir/wildfires-in-canada-19502021.