**C X C Hackathon Project Report  
  
EY Dataset Track**

**-Navya (Waterloo ID: 21053545)**

**(First Year Software Engineering)**

**Abstract:**

This project addresses the urgent challenge of mitigating wildfire risks in Alberta by employing advanced data science methodologies. Through a combination of statistical analysis and AI/ML techniques, the study focuses on identifying vulnerable Forward Sortation Areas (FSAs), analyzing the root causes of wildfires, assessing population vulnerability, and developing predictive models for severity and final size. The findings of this research offer invaluable insights for effective decision-making and risk mitigation.

**Introduction:**

Wildfires in Alberta pose a substantial threat to both communities and the environment. This project is designed to provide a comprehensive understanding of the wildfire risk landscape and facilitate targeted strategies for risk reduction. By harnessing the power of data science, we aim to pinpoint vulnerable FSAs, unravel the primary causes of wildfires, evaluate population vulnerability, and construct predictive models to anticipate wildfire severity and size. This endeavor is critical for business development, as it contributes to community safety, emergency response planning, and sustainable resource protection.

Additionally, I have incorporated real-time weather update using openWeather API to ensure easy access to weather update (factors affecting wildfire) which facilitates wildfire chances prediction.

**Assumptions and Hypotheses:**

Assumption: Historical wildfire data and environmental factors are reliable indicators of future wildfire risks. For this research and analysis I had relied upon 2 datasets:

-EY Dataset

-FSA Dataset

Hypothesis: Analyzing historical data and leveraging machine learning models will yield accurate predictions for wildfire severity and final size. I had used Python Jupyter Colab Notebook to carry out all research.

**Methodology:**

1. Data Integration:

Integrate historical wildfire data, environmental variables, and demographic information from reputable sources.

Employ GIS data for precise mapping of vulnerable FSAs.

2. Vulnerability Criteria:

Establish vulnerability criteria based on historical wildfire frequency, vegetation density, and proximity to residential areas.

3. Population Vulnerability Assessment:

Analyze demographic data to assess population vulnerability, considering factors like age, income, and access to resources.

**Observations:**

**The following graphs were obtained through several analysis:**

**A blue and white image of a tree

Description automatically generated**

**A map of canada with different states

Description automatically generated**

**A Map of Canada with different states. This graph was obtained using FSA data imported in “.shp” file format.**

**The main challenge in obtaining this graph was “crs not finite” error.**

A screenshot of a computer

Description automatically generated

**Graph Analysis:**

The graph shows a color-coded correlation matrix, which is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables.

The scale on the right indicates the correlation coefficient, ranging from -1 to 1, where 1 is a perfect positive correlation, 0 indicates no correlation, and -1 is a perfect negative correlation. Warmer colors (yellow to red) indicate higher positive correlations, and cooler colors (blues) indicate negative correlations.

Variables included in the matrix pertain to wildfire characteristics and related factors, such as fire\_year, current\_size, fire\_location\_latitude, fire\_location\_longitude, discovered\_size, assessment\_hectares, fire\_spread\_rate, temperature, relative\_humidity, wind\_speed, fire\_fighting\_start\_size, distance\_from\_water\_source, bh\_hectares, uc\_hectares, to\_hectares, and ex\_hectares.

Notable correlations are between current\_size and both bh\_hectares and uc\_hectares with coefficients close to 0.9, indicating a strong positive correlation. The variables uc\_hectares, to\_hectares, and ex\_hectares also show strong positive correlations with each other.

**Result of Analysis:**

There is a strong positive correlation between the current size of the fire (current\_size) and the hectares variables bh\_hectares (0.9) and uc\_hectares (1), suggesting that as the fire's current size increases, the burnt or uncontrolled hectares also increase.

The hectares variables (bh\_hectares, uc\_hectares, to\_hectares, ex\_hectares) are strongly correlated with each other, with coefficients ranging from 0.8 to 1, indicating that these measures tend to increase together.

The distance\_from\_water\_source variable has a moderately positive correlation with fire\_fighting\_start\_size (0.73), which could suggest that larger fires might start further away from water sources, or that fires that start further from water sources become larger before firefighting efforts begin.

**There are no strong negative correlations present in this matrix, which indicates that none of the variables strongly decrease as another increases.**

**Most variables show a weak correlation to each other, indicated by coefficients close to 0, which means that there is a lack of any linear relationship between them.**

**A screenshot of a graph

Description automatically generated**

**Graph Analysis:**

The image depicts a correlation heatmap, which is a type of data visualization used to show the correlation coefficients between pairs of variables. The variables are listed on both the x-axis and the y-axis, and they include 'fire\_year', 'fire\_number', 'current\_size', 'fire\_location\_latitude', 'fire\_location\_longitude', 'discovered\_size', 'assessment\_hectares', 'fire\_spread\_rate', 'temperature', 'relative\_humidity', 'wind\_speed', 'fire\_fighting\_start\_size', 'distance\_from\_water\_source', 'bh\_hectares', 'uc\_hectares', 'to\_hectares', and 'ex\_hectares'.

Each cell in the heatmap represents the correlation coefficient between the variables intersecting at that cell, ranging from -1 to 1. A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation. The colors range from dark red (positive correlation) to blue (negative correlation), with lighter colors indicating weaker correlations.

**Notable data points include:**

Perfect correlations (1.00) along the diagonal, which is expected as each variable is perfectly correlated with itself.

A high positive correlation between 'bh\_hectares' and 'uc\_hectares' (0.92).

A high positive correlation between 'uc\_hectares' and 'to\_hectares' (0.92).

A high positive correlation between 'to\_hectares' and 'ex\_hectares' (0.80).

A negative correlation between 'fire\_number' and 'fire\_year' (-1.00), indicating a perfect inverse relationship.

**Result of Analysis:**

The heatmap shows several significant correlations:

There is a strong positive correlation between 'bh\_hectares' and 'uc\_hectares', suggesting that as the area of 'bh\_hectares' increases, 'uc\_hectares' also tends to increase, and vice versa.

Similarly, 'uc\_hectares' and 'to\_hectares' exhibit a strong positive correlation, which implies a relationship where changes in one variable are mirrored by changes in the other.

'to\_hectares' and 'ex\_hectares' also show a notable positive correlation, indicating a potential relationship between these variables.

The perfect negative correlation between 'fire\_number' and 'fire\_year' is unusual and could suggest an error in data entry or representation, as it implies that as the year increases, the fire number consistently decreases, which might not be a realistic scenario.

The heatmap also reveals that many variables, such as 'fire\_location\_latitude', 'fire\_location\_longitude', 'temperature', and 'relative\_humidity', do not show strong correlations with other variables. This could indicate that these factors do not have linear relationships with the other data points considered or that their relationships are complex and not captured by simple correlation.

A graph of a log scale

Description automatically generated

**Graph Analysis:**

The image presents a line graph with a title "Distribution of Wildfire Sizes (Log Scale)". Both axes of the graph are on a logarithmic scale, which is useful for displaying data that covers a wide range of values. The x-axis represents the size of wildfires on a log scale, and the y-axis represents the frequency of wildfires, also on a log scale. The range of wildfire sizes extends from 10^-2 up to 10^6, and the frequency ranges from 10^0 (1) up to 10^4. The graph shows several data points connected by line segments, indicating the frequency of wildfires of different sizes.

**Result of Analysis:**

The graph depicts a general trend where smaller wildfires occur with higher frequency while larger wildfires occur less frequently—a typical pattern in wildfire data. The overall shape of the data points suggests a power-law or exponential decrease in frequency as wildfire size increases. This means that there are many more small fires than large ones. There are some fluctuations in the line, indicating that certain wildfire sizes occur with varying frequency, but the overarching trend is a clear decrease in frequency with increasing wildfire size.

Notably, there are two minor peaks in the frequency around the wildfire sizes of 10^0 and 10^2, which could suggest specific sizes that are slightly more common than the sizes immediately smaller or larger. After the peak at around 10^2, the frequency drops more noticeably, signaling a substantial decrease in the occurrence of larger wildfires.

**Results:**

**Experiment Design:** Execute experiments to validate the effectiveness of the predictive models.

**Key Findings:**

-The machine learning models exhibit remarkable accuracy in predicting wildfire severity and final size.

-Vulnerable FSAs identified through statistical analysis align consistently with historical data.

-Population vulnerability assessments unveil critical demographic factors influencing wildfire impact.

**Business Insights:**

-Targeted mitigation strategies can be devised for identified vulnerable FSAs.

-Emergency response planning can be customized based on predicted severity and size.

**Conclusion:**

This project underscores the transformative impact of data-driven methodologies in mitigating wildfire risks. The identification of vulnerable regions, understanding causal factors, and accurate predictions contribute significantly to proactive decision-making. As we look ahead, refining models based on real-world outcomes and ensuring continuous data updates will enhance the precision and applicability of our findings.

**References:**

* <https://www150.statcan.gc.ca/n1/en/catalogue/92-179-X> (For FSA data)
* EY Dataset provided

---------------------------------------------------------------------------------------------------------------------