
# Application and Evaluation of Forest Fire Size Prediction Model Based on Machine Learning Algorithms

## Final Report

By Ni Zikun

BEMM466J - Business Project

Professor Salimeh Pour Mohammad & Beth Kewell

Jan 2022

# Contents

## 1. Introduction

With global warming and human activities, forest fires are growing increasingly extreme and widespread, posing a huge threat to the environment, property, and human health. In 2020, due to forest fires, tropical areas around the world lost 122 thousand square kilometres of the forest, a 12% increase over the previous year (Weisse & Goldman, 2021). Within humid tropical primary forests, which are essential for carbon sequestration and biodiversity, 42 thousand $km^2$ of forest have been destroyed, equivalent to the size of the Netherlands. The greenhouse gases resulting from these forest fires are equal to the yearly emissions of 570 million automobiles, which is more than double the number of automobiles in the US.

In order to deal with the increasing damage of forest fires, governments around the world have invested a great deal of capital and talent into forest fire study. Forest fire size prediction is critical for firefighting in forest fire research. The size of forest fires affects the amount of pollution they produce, how many people need to be evacuated, and how many rescuers and supplies are required. Therefore, estimating the final size of forest fires early on can assist fire departments in formulating effective rescue plans to reduce losses. Fortunately, the development of big data and artificial intelligence has made it possible to predict the size of a forest fire. According to the US Wildfire Classification System, fire size can be divided into seven levels. Therefore, predicting the forest fire size is essentially a classification task. The prediction model with environmental data as input and fire size as output can be realised using machine learning classification techniques. At present, researchers mainly focus on predicting forest fire occurrence probability through environmental factors, and relatively few studies on the prediction of forest fire size. In addition, some research on fire size prediction only considers weather factors, resulting in low prediction accuracy. In fact, an increasing number of research found that the fuel moisture factors have a critical effect on the size of a forest fire.

The objective of this research is to establish a model that can accurately estimate the magnitude of forest fires through various machine learning algorithms and fire-related factors. In addition to the weather factor, the fuel moisture was introduced into the prediction model in this project, and features were appropriately reduced through feature selection to improve the model operation efficiency and prediction accuracy. The historical fire data from Alberta, Canada, was collected as experimental dataset in this project. Related feature datasets, including historical ignition point fuel moisture dataset, weather dataset and elevation dataset were gathered from the Canadian Wildland Fire Information System Datamart, Canadian Weather site, and the Google Elevation Application Programming Interface, respectively. The research steps of this project are shown in Figure 1.

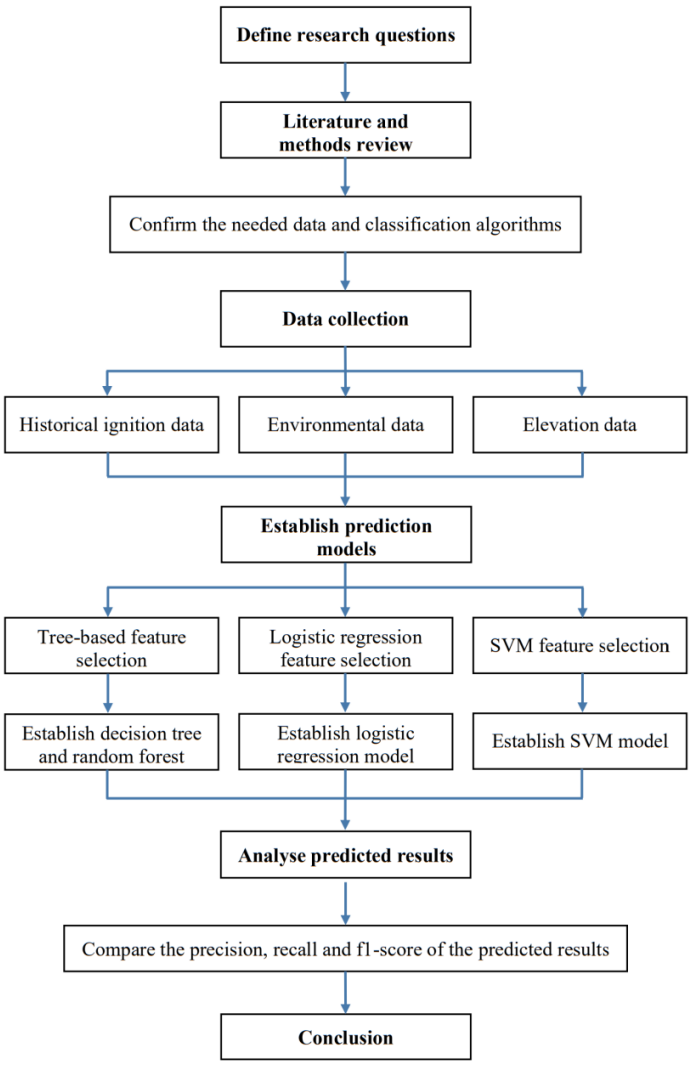


**Figure 1.** Research steps

First, this project will conduct a literature and method review to identify the related factors affecting forest fire size and effective classification algorithms. Then, the historical ignition point data and related factors will be collected through public databases. Next, key features will be selected based on several classification algorithms. After that, forest fire predictive models will be developed using these classification algorithms and key features. Finally, the optimal predictive model will be determined by evaluating the models' predictive performance.

## 2. Context and Background

Forest is the main body of the terrestrial ecological system and the indispensable natural resource for human beings. It is known as "the lungs of the earth" and is essential for maintaining balance of nature and optimising the ecological environment. However, forest resources have a severe threat, that is forest fires.

Forest fires destroy timber, wildlife habitats and watersheds, and cause forest ecosystems to lose balance. In addition, forest fires pose a risk to humans and their property, especially in areas where forest region and developed areas collide (US Geological Survey, 2006). The forest fire will increase the incidence of flooding, mudflows, and landslides. It will also cause severe health problems due to smoke and other emissions. Data from Forest Research (2021) show that around 1000 hectares of woodland in the UK are burnt every year, and the annual cost of fighting forest fires is 55 million pounds. According to the Insurance Information Institute (2021), the catastrophic and sustained wildfire on the West Coast of America combusted over 16,187 $km^2$ in 2020, resulting in 2.12 billion GBP in property damage. The dry weather in Indonesia resulted in wildfires destroying over 1 million hectares in six Indonesian provinces, and bushfires in Australia burned 186,155 $km^2$ and displaced billions of wildlife during the same year (Climate Reality Project, 2020).

Faced with such devastation, humans are increasingly focused on forest fire prediction and prevention. Over the last few years, the forest fire size prediction has been an essential field in forest fire research. Predicting forest fire size by intelligent technology can help fire departments to estimate losses and formulate effective rescue plans. Therefore, this research can help to decrease the amount of damage caused by wildfire and will contribute significantly to forest resource protection and the ecological balance maintenance.

Machine learning techniques are developing rapidly, making it possible to estimate the size of wildfires. It is basically a classification issue when it comes to predicting the size of fires.

Classification algorithms such as Random Forest, Logistic Regression, Decision Tree, and SVM in machine learning can help solve this problem. These algorithms can learn and model by large amounts of forest fire data in supervised or unsupervised methods. After training, the model will be able to make accurate predictions based on data features. Therefore, this project will realise the size prediction of forest fires using the machine learning technology.

## 3. Research Objective and Questions

### *3.1. Research Objective*

The objective of this project is to establish a model which can correctly predict the size of wildfires. The model will be able to estimate the forest fire size according to the input fire size related factors. This project will train forest fire prediction models based on different algorithms and key features. Then, the performance of Precision, Recall and F1-Score of models will be evaluated to determine the optimal prediction model. The predicted size of forest fires can support fire departments in developing effective rescue measures and reducing losses caused by forest fires.

### *3.2. Research Questions*

To achieve the aim, this project includes six research questions.

(1) What are the main influence factors in wildfire size?

(2) Which algorithms are suitable for fire size prediction?

(3) What are the key features of fire-size predictive models based on tree-based, Logistic Regression, and Support Vector Machine algorithms, respectively?

(4) How effective are fire-size predictive models based on the key features?

(5) Which machine learning model has the highest predictive evaluation?

## 4. Literature Review

### *4.1. Global Trends in Forest Fire*

The global climate is becoming progressively drier and warmer (Allen, 2010). According to IPCC, the global warming results from intensified greenhouse effects caused by human activities. The global average radiation intensity due to greenhouse gases will continue to rise throughout the 21st century, according to IPCC's emission report (Nakicenovic et al., 2000). The global forest fires occurrence frequency, burning area, and fire intensity have all increased significantly as a result of global warming (Batllori, 2013). Wibbenmeyer and McDarris (2021) found that the average number of wildfires in the US has decreased by 780 per year over the last three decades, however, the overall area burnt has climbed by 192,000 acres each year. From 1970 to 2000, the average annual area combusted by large wildfires increased by 1200 %. In America, over 9 million acres of forest were burned in 2006, and around 1.2 million acres were burned in 2020 (Wibbenmeyer & McDarris, 2021). As of July 2021, 15 of the 20 most catastrophic forest fires in California history have happened after 2015. In recent years, large forest fires have ravaged portions of northern and western Europe that have never been burnt before (European Environment Agency, 2021). In 2017, forest fires in the Mediterranean nations of Europe burned the second-largest area on record, particularly in Portugal, where unprecedented forest fire destroyed 1.3 million acres (Turco et al., 2016). Meanwhile, more European nations, especially Central and Northern Europe, saw significant wildfires in 2018 compared to the past. In 2018, Sweden had its worst fire weather, necessitating international firefighting help via the European Civil Protection Mechanism (European Commission, 2018). According to Canada's National Forestry Database (2020), between 2009 and 2019, the yearly forest area burned in Canada was very unpredictable, with no discernible trend. With nearly 11 million acres burned in 2014, it was the year with the most area scorched. With roughly

800,000 acres of land burned in 2009, it was the year with the fewest area scorched. Following two record-breaking forest fire seasons in 2017 and 2018, the area burned in British Columbia in 2019 was lower than the long-term norm. In 2019, 4,000 forest fires burnt over 4.4 million acres of forest in Canada, which is 30 percent less than the 25-year average (Natural Resources Canada, 2020).

The rising threat of forest fire not only harms the forest system but also seriously affects human habitation and survival. Forest fire smoke has increased significantly in the United States since 2000, and by now, forest fires expose about 25 percent of Americans to the harmful PM2.5. In 2020, nearly 18,000 structures were destroyed in the United States forest fires, including nearly 10,000 homes (Wibbenmeyer & McDarris, 2021). In Canada, Northern Alberta experienced 4 times the 10-year average burnt area in 2019, causing the evacuation of Wabasca and High-Level communities. The area burnt in Yukon was double the 10-year normal due to a dry winter and spring. Several evacuations, road closures, and extensive smoke persisted unusually late into September. In the same year, as two distinct wildfires blazed nearby, the Pikangikum First Nation in Ontario was totally evacuated in July (Natural Resources Canada, 2020). In 2010, a series of forest fires occurred in Russia, with a total of 813 forest fires, and Russia invested more than 240,000 people, 60 aircraft and 120,000 fire-fighting equipment in fighting the fires (Sano et al., 2011).

### *4.2. The Beginning of Forest Fires Forecasting*

Forest fires forecasting is the process of analysing and forecasting the difficulty of fire control, the possibility of fire occurrence, and fire behaviour indicators using a variety of features such as weather, terrain, fire source, dry and wet degree of fuel, and fuel type. The study of forest fire forecasting began when Coert Dubois (1914) published his research of systematic forest fire protection. Coert Dubois discussed the concept and meteorological elements of forest fire

forecasting but did not put forward the method of fire prevention measurement or fire forecasting. Since then, forest fire forecasting research has developed rapidly. During the Tsarist era, the Former Soviet Union began conducting experiments on juniper branches to predict the possibility of forest fires. Wright (1937) predicted forest fires by relative humidity and proposed that forest fires were possible when relative humidity was lower than 50%. Gisborne (1936) conducted research on accurately predicting fire risks based on meteorological factors and proposed the multi-factor method of forest fire risk prediction, which was the first forest fire risk prediction system in America.

### *4.3. Fire Danger Rating System*

From the 1950s to the 1960s, an increasing number of countries investigated forest fire forecasting. Among these countries, America, Canada, and Australia had more advanced research progress.

In 1972, the United States released the National Fire Danger Rating System (Deeming et al., 1977). This system is a physical model developed based on combustion principles and laboratory tests. The constants and parameters used in the model reflect the relationship between fuel, weather, terrain, and other factors. The National Fire Rating System consists of three components (Ignition Component, Spread Component, and Energy Release Component) and three indexes (Lightning Fire Occurrence Index, Human-caused Risk Scaling Index, and Burning Index). The Fire Load Index (FLI) is a combination of these three indexes that represents the challenge of putting out the fire and the firefighting workload. Therefore, the NFDRS is a comprehensive system capable of forecasting not only fire risk, but also forest fire occurrence and behaviour. As technology advances, more and more new technologies are being applied to the NFDRS. For example, the Surface Observations Gridding System (SOGS) was used to process meteorological data (Jolly et al., 2004), the Nelson model can measure the

moisture content of the fuel (Nelson, 2000), and the Normalized Difference Vegetation Index (NDVI) can measure the vegetation condition (Burgan et al., 1996).

Canada released the Forest Fire Weather Index System in 1974, which became the country's national wildfire forecasting system (Van Wagner, 1987). Canada released the Canadian Forest Fire Danger Rating System (CFFDRS) in 1987. CFFDRS is a more advanced system that includes three parts: the Fire Weather Index (FWI) System, the Fire Behaviour Prediction (FBP) System, and the Fire Occurrence Prediction System. The FWI is an empirical system for predicting fire risk based on meteorological factors, fuel moisture content calculations, and small-field ignition tests. The system takes 12-14 hr relative humidity, temperature, windspeed and the 24-hr rainfall as the basic input variables. Fine Fuel Moisture Code, Duff Moisture Code, and Drought Code in the system indicate water concentration in different fuels. Initial Spread Index, Buildup Index, and Fire Weather Index are three output indexes that reflect spread speed, energy release rate, and fire behaviour potential (McAlpine, 1990).

Forest fire forecasting in Australia began in the 1960s. McArthur (1958) developed the Fire Danger Tables according to the prediction of fire-fighting difficulties under various weather conditions using the spread rate of standard dry wildfires. In 1973, McArthur released the Forest Fire Danger Meter (FFDM). Based on the combination of ambient temp, moisture content, speed of gust, and drought impacts, the FFDM can calculate the fire risk index that takes into account the frequency, spread speed, severity, and suppression difficulty of the fire. Since then, the Forest Fire Danger Meter has been successfully adopted and deployed by regional Fire Department and the Bureau of Meteorology in Australia. The SiroFire System, which is presently used in Australia, is a PC-based decision-making tool. The SiroFire System is based on the FFDM developed by McArthur and the Grassland Fire Danger Meter developed by Commonwealth Scientific and Industrial Research Organisation (CSIRO). This system uses the Fire Danger Meter algorithm to estimate the spread of wildfires and draw the fire boundary

maps according to the input from fire controllers on temperature, moisture content, gust speed, fuel, and slope direction (San-Miguel-Ayanz et al., 2003).

### *4.4. The Influence of Meteorological Factors*

#### 4.4.1. Temperature

Temperature is usually measured by the meteorological station at 2 meters above the ground. The average kinetic energy of molecules in the atmosphere increases when the temperature rises and decreases when the temperature drops. Temperature is mainly affected by altitude and long-wave radiation. The higher the altitude is, the worse the insulation of the atmosphere is, so that the temperature at the high altitude in the same area is lower than that at the low altitude, with a specific rate of 0.6 °C temperature drop for every 100 meters of elevation rise. As for the impact on forest fire risk, temperature indirectly changes the moisture concentration of fuel mainly by changing the relative humidity. With the temperature increasing, the decrease of the moisture content of surface fuel will greatly reduce the energy required for igniting fuel, leading to the increase of forest fire probability. Therefore, the temperature of a day has always been an important parameter of wildfire prediction models. From the statistics of wildfires, when the highest temperature reaches 12 °C, it may lead to forest fires, and most forest fires occur in temperatures above 15 °C (Fried et al., 2004).

#### 4.4.2. Relative Humidity

Relative humidity refers to the air dryness. Because relative humidity is mainly influenced by temperature, it has maximum and minimum values throughout the day. Relative humidity will affect the fuel's moisture content, which has a close relationship with forest fire occurrence. The balance between relative humidity and fuel moisture makes fuel gradually lose water when it is higher than relative humidity and absorb water when lower than relative humidity. The

humidity level is also affected by fuel size. The smaller the diameter of the fuel, the shorter the response time to relative humidity. Wright et al. (1982) found when the humidity in the air falls below 50%, forest fires are more probable to erupt. Flannigan (2001) used relative humidity to predict forest fires. Therefore, relative humidity is the main factor affecting forest fire risk, which can directly determine the probability of forest fire.

#### 4.4.3. Wind

Wind is the horizontal movement of air, and wind speed reflects the speed of this movement. The metre per second (m/s) is a common unit of measurement for speed. The size of wind speed can influence the rate of fuel moisture loss, thus changing the probability of forest fire occurrence (Wagner, 1979). Byram & Jemison (1943) found that wind can speed up the combustion process by facilitating air circulation and providing a continuous supply of oxygen for forest fire. During forest fire, the wind flow will produce heat convection to dry the fuel near the fire and promote forest fire spread. Therefore, wind speed is also an essential factor in forest fire prediction models.

#### 4.4.4. Precipitation

Precipitation includes vertical precipitation falling from clouds to the ground and horizontal precipitation generated in fog, frost, dew. Precipitation is the sum of these two parts, and the precipitation unit is $mm$. Precipitation is mainly influenced by climate and geography. Like relative humidity, precipitation mainly affects the process of fuel moisture loss and absorption (Wotton, 2007). In addition, precipitation can affect ongoing forest fires by lowering the temperature of the fire and depriving them of oxygen. Viegas et al. (2001) discovered a negative association between 24-hour precip and the quantity of forest fires. Therefore, precipitation is an important factor affecting forest fires.

### *4.5. The Influence of Fuel Moisture Factors*

Impacts of fuel moisture concentration on the combustion process is an essential theoretical basis for studying the relation between fuel moisture and forest fire. The moisture content directly participates in combustion processes and has a cooling effect on the combustion of wood materials. The higher the fuel moisture, the higher its heat capacity, then more heat is required to evaporate this moisture. If the moisture level is high enough, the combustion process can be interrupted. This moisture content is called moisture of extinction (MOE). The MOE of fine fuels is usually 30% to 35% (Catchpole, 1993).

Fuel is usually classified into two categories: living fuel and dead fuel. Live fuel is part of the plant organism. During the growing season, live fuel dominates in forests. Its water concentration is relatively stable under the physiological effect of plant water balance, around 120%, and is usually higher than that of dead fuel. Compared with the dead fuels' moisture level, the live fuels' moisture level does not change significantly with the change of meteorological conditions in a short time. Remote sensing images can predict the forest live fuel's moisture concentration in the growing season to predict the potential spreading speed and fire intensity of forest fire (Chuvieco, 2004). Dead fuels, such as fallen branches, litter, and foliage, are mostly on the forest floor. The water concentration of dead fuels has a significant impact on not only the spread speeding and intensity of a fire, but also the wildfire igniting probability. In addition, dead fuels' moisture is more susceptible to dynamic changes in meteorological factors.

Many forest fire forecasting models take dead fuel moisture into account. When Marsden-Smedley (2001) developed a fire model to estimate the fire danger of Tasmanian Buttongrass Moorlands, he discovered that dead fuel moisture had a significant impact on fire spreading speed. The indoor fire experiments conducted by Plucinski et al. (2010) showed that the amount of litter, vegetation moisture, shrub density, wind speed, and dead fuel amount were major

factors influencing the spread of the shrubland fire. Viegas et al. (1992) discovered a significant negative association between dead fuels' moisture concentration and wildfire incidence in Portugal. Renkin & Despain (1992) discovered that the quantity and spread speed of forest fires were related to dead fuel's water content in Yellowstone National Park. Nash and Johnson (1996) discovered similar results in Canada's Central Provinces. Ray et al. (2005) discovered the moisture concentration of litters had a noticeable impact on the occurrence of wildfires in the Brazilian Amazon Basin. Chuvieco et al. (2009) established a forest fire prediction model that reflected the relationship between dead fuel's water concentration and forest fire behaviour using Logistic Regression algorithm. Therefore, dead fuels' water content is essential to wildfire prediction. The fuels' moisture concentration mentioned below is the dead fuels' moisture concentration.

#### 4.5.1. Fuel Moisture Codes

Three fuel moisture codes, Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC) and Drought Code (DC), respectively, indicate the water concentration of forest litters under the shade of a forest canopy, decayed plant residues beneath the litters, and the deep soil. The most significant difference between these three layers is their different decomposition stages. They are subdivisions of humus or organic layer based on soil science, representing litter layer, fermentation layer and Humus layer in humus or organic layer, respectively. Fuel moisture codes are not only used as the intermediate variable in computation processes of the Canadian Fire Weather Index but also the core model for moisture content prediction in the US BEHAVE system. Research by Flannigan (2016) based on global climate change models shows that for every 1°C increase in global warming, FFMC, DMC and DC will increase by at least 15%, 10% and 5%, respectively, and the extreme fire weather occurrence will greatly increase. When FFMC is less than 75, the moisture content of fine fuel is close to 30%, and the fine fuel will

not ignite and spread. When DMC is less than 20, the fermentation layer will not participate in the combustion even if the litter layer is ignited (Stocks et al., 1989). National researchers and forest fire managers are increasingly recognising the theory and form of the fuel moisture codes. It has been applied in Canada, some states of the United States, Australia, New Zealand, Spain, Portugal, Chile, Indonesia and other fire-prone countries and regions. In addition, fuel moisture codes are also used in canopy fire prediction, surface fuel consumption prediction, fire size prediction, forest fire carbon emissions and many other aspects of research (Amiro et al., 2001).

4.5.2. Fine Fuel Moisture Code (FFMC)

FFMC refers to fine fuels' water concentration with a ground thickness of 1.2 $cm$ and an areal density of 0.25 $kg/m^2$. The moisture content ranges from 0 to 250%. The conversion formula between fine fuel moisture content and FFMC is as follows.

$$M = 147.2(\frac{101 - FFMC}{59.5 + FFMC})$$

In the formula, $M$ means the predicted value of the fine fuel's water concentration. When M equals to 0, the FFMC value is equal to 101. As a result, the higher the FFMC value, the greater the danger of fire (Van Wagner, 1987). The weather in the forest usually changes regularly. Since the fine fuel moisture content is mainly affected by meteorological factors, the fine fuel moisture will change regularly with the forest weather. Under the same meteorological conditions, the rate of water loss and water absorption of fine fuel is the largest compared with other fuel types. Therefore, in order to study the dynamic law of fine fuel moisture content on different time scales, Van Wagner (1987) successively proposed the FFMC of daily calculation and hourly calculation to meet the moisture content prediction requirements of different precision and scale. Lawson (1996) proposed an hourly calculation model for FFMC based on Van Wagner's research, which can compute fine fuel's moisture concentration in the Canada fire behaviour prediction system.

4.5.3. Duff Moisture Code (DMC)

DMC is used to indicate the water concentration of a duff fuel with the thickness of 5-10 $cm$ and the areal density of 5 $kg/m^2$, consisting of organic material that has begun to decompose but whose form and origin can still be identified. DMC ranges from 0 to infinity. The larger the DMC is, the lower the moisture content of duff fuel is. The conversion formula between duff fuel moisture content and DMC is as follows.

$$M = 20 + e^{5.6348 - \frac{DMC}{43.43}}$$

In the formula, M represents the predicted duff fuels' water content. DMC has been adopted to estimate lightning fires' incidence in Canada. DMC equal to 20 is defined as the threshold at which duff fuel can be ignited by lightning in Alberta and Ontario (Wotton, 2005). When DMC is equal to 0, the duff fuel moisture content is 300%. Although the value of DMC has no upper limit, DMC greater than 150 is rarely seen in practice. Generally, DMC greater than 27 will be defined as a high fire risk. Because Duff fuel is covered by fine fuel, DMC is usually only affected by rainfall, temperature, and humidity level. Rainfall with less than 1.5 mm of total rainfall in 24 hours usually have no influence on DMC value. Because the water loss and water absorption rate of duff fuel are lower than that of fine fuel, the change of DMC on the hourly scale is minimal. Therefore, only the daily calculation model of DMC is established, without considering the changes in hourly calculation.

4.5.4. Drought Code (DC)

DC indicates the deep soil's water concentration with the ground thickness of 18 $cm$ and the areal density of 25 $kg/m^2$. The moisture content ranges from 0 to 400%. The conversion formula between fine fuel moisture content and DC is as follows.

$$M = \frac{800}{e^{\frac{DC}{400}}}$$

In the formula, M represents the predicted value of the deep soil moisture content. DC is usually used to indicate long-term seasonal drought conditions, and the value of DC is closely related to groundwater level (Wotton, 2001). Because the moisture content of deep soil rarely participates in the ignition and spread, DC has little influence on the forest fire.

### *4.6. Summary*

Forest fires are devastating to the environment, human health, and property, and they are growing increasingly dangerous because of global warming. Hence, many countries have invested heavily in forest fire prediction research. According to the literature review, the main factors affecting forest fire are temperature, wind speed, relative humidity, precip, fine fuel moisture code (FFMC), duff moisture code (DMC) and drought code (DC). This paper will collect the above fire-related factors and historical wildfire dataset and build the forest fire size prediction model according to these data.

## 5. Overview of Methods and Choice of Methods

### *5.1. Forest Fire Prediction Methods Overview*

Forest fire prediction methods are usually based on mathematics and statistics in order to develop predictive models of forest fires and related factors like climate, terrain, and plants. Because forest fires' size can be divided into several tiers, forecasting a wildfire size is basically the classification problem. Machine learning classification algorithms like Logistic Regression and Decision Tree are capable of predicting the forest fires' size (Harrington, 2012).

Logistic Regression (LR) is a common method in forest fire prediction research. Prasad et al. (2008) adopted satellite remote sensing technique and LR algorithm to analyse effects of terrain, plants, temperature, and mankind behaviours to forest fire occurrence and developed a LR program which can forecast wildfires' incidence. Vega-García et al. (2010) studied the association between terrain heterogeneity and forest fires in the Mediterranean area through LR model. Chang (2013) predicted the probability of forest fire in Heilongjiang province by terrain, vegetation type and weather using the Logistic Regression. Wotton et al. (2005) predicted lightning-fires' incidence in Ontario using weather and fuel moisture content.

The geographically weighted regression (GWR) is also being adopted to the research of wildfire occurrence prediction models. Geographically weighted regression adds spatial dataset to the prediction model and weighted spatial matrices to regression coefficients. Coefficient's value of every spot changes as the spatial position changes. GWR is able to reduce effects from spatial non-stationarity of variables to the model (Fotheringham, 1996). Rodrigues et al. (2014) applied GWR to the logistic model to analyse the influence of the social economy, traffic, land type and other factors on the forest fires occurrence in Spain and drew the distribution map of forest fires according to the Logistic GWR model. The findings suggest that the Logistic GWR model performs better than the classical LR method. Nevertheless, GWR is only able to analyse

quantitative variables, not able to analyse categorical variables, resulting in a certain limitation in variable selection while using GWR to study forest fire factors.

Decision Tree (DT) is also an effective classification method. In essence, DT is a tree-structured classifier. Nodes in the decision tree represent sample feature tests, then the algorithm would classify test samples based on their unique features. The training samples will be systematically segmented to smaller subgroups when decision trees are established, until all feature tests are finished. ID3, C4.5, CART, and CHAID are some of the most widely used DT algorithms. Quinlan created the ID3 method in 1979, and it leverages the idea of entropy in information theory to evaluate whether or not it should create nodes through the information gain. C4.5 method addresses ID3's flaw of only selecting attributes with larger coefficient (Coffield et al, 2019). The DT algorithm can handle enormous data sets and has the benefits of speed and simplicity. Coffield et al. (2019) classified wildfires' size into three categories: huge, normal, and small, then utilised dataset of wildfires in Alaska to create a wildfire size forecast model using the DT algorithm, which has a 50.4 percent accuracy. Lang Qin et al. (2021) developed a wildfire size forecast model using the DT algorithm according to the forest fire dataset and meteorological dataset in America, which has a 67 percent accuracy.

Random Forest (RF) algorithm is also a method widely used in forest fire research. Breiman Leo and Adele Cutler proposed the Random Forest algorithm in a research article published in 2001. The Random Forest algorithm is capable of classification, grouping, and regression. It realises categorisation by developing several decision trees and analysing their votes outcomes by Bootstrap sampling method (Collins et al, 2020). The RF algorithm could efficiently process many independent factors and automatically choose essential factors. Moreover, the Random Forest algorithm will not be influenced by the multicollinearity among factors. Therefore, the RF algorithm can flexibly evaluate the complex relationship among variables. Archibald et al. (2009) utilised the RF algorithm to study influences from human activities and climate factors

on the burned area in South Africa and the importance of each variable in different regions. Oliveira et al. (2012) used the RF algorithm to analyse the impact of climate, social economy, population, and other factors to wildfires' occurrence in the Mediterranean. They compared it with the traditional multiple linear regression method to study the applicability of the two modelling methods in estimating wildfires incidence in Europe. Their findings suggest that the prediction performance based on RF algorithm is higher than that based on multiple linear regression algorithm. Nevertheless, the RF algorithm has the disadvantage of being unable to compute the regression coefficient and confidence interval.

In addition, some other methods are adopted to study the wildfire-related factors and prediction models. Miranda et al. (2012) analysed wildfires' temporal and spatial distribution from northern Wisconsin and used linear regression to quantify the impact of weather. The results show that drought and population are key factors in wildfires' occurrence and spread, and the inhabitant density is non-linear correlated with wildfires' incidence. McCOY (2005) used multiple linear regression to analyse the effects of average temperature, total precip, relative humidity, and wind speed to fire occurrence, fire size, and severity. Podur et al. (2010) predicted forest fires' size in southern Ontario using Poisson regression.

### *5.2. Choice of Methods*

According to the above methods review, the prediction performance of existing forest fire size prediction models is low, and usually only meteorological factors are considered. Hence, this project added climate factors and fuel moisture codes into modelling to achieve higher predictive accuracy. The literature review shows that Logistic Regression algorithm is commonly adopted in wildfires forecast, and tree-based algorithms will not be affected by multicollinearity between variables. In addition, the Support Vector Machine is also an effective classification algorithm but is rarely employed in wildfire area forecast. Therefore, in

this project, the Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression algorithm will be used to establish forest fire size prediction models, which take the key forest fire-related factors as input and the fire size as output.

### *5.3. Modelling Procedures*

(1) In the data collection step, this project will collect the dataset of historical forest fires in Alberta from the Canada National Natural Resources Datamart. Datasets of historical ignition points' weather, fuel moisture codes and elevation will be collected from the Canada Historical Climate website, Canadian Wildland Fire Information System Datamart, and Google Elevation Application Programming Interface, respectively. The web crawler technology will be utilised in this project to acquire a large quantity of historical weather data for each ignition site.

(2) In the data pre-processing step, this project will merge the four datasets collected in the previous step and remove the data containing the null value. The forest fire size will be divided into four tiers based on the categorization system of wildfires from America National Wildfire Coordinating Group. And then all features' value will be normalised to the range 0 to 1.

(3) In the feature selection step, forest fire-related features will be selected for modelling by comparing the correlation coefficients between forest fires size and these features based on Tree-based algorithm, Support Vector Machine, and Logistic Regression, respectively.

(4) In the modelling and evaluation step, models will be trained in Python using Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression algorithm. The optimal forest fire size prediction model will be determined by evaluating and comparing the performance of precise, recall and F1-score of the models' predictions.

## 6. Findings and Discussions

### *6.1. Data Collection and Pre-Processing*

#### 6.1.1. The Historical Forest Fire Data

The historical forest fire dataset was downloaded from Canada's Natural Resources Datamart (2020). The dataset contains 413,150 historical forest fire data in Canada from 1950 to 2019, including dimensions such as province, fire number, latitude, longitude, date, combustion area, and type. Table 1 shows samples of historical forest fire dataset.

**Table 1.** Samples of historical forest fire dataset

| Prov | Fire Id | Lat | Lng | Attk Date | End Date | Size ha | Fire Type |
|---|---|---|---|---|---|---|---|
| AB | HWF132 | 58.003 | -116.685 | 2011/9/20 | 2011/9/23 | 0.22 | Forest |
| AB | P04006 | 53.196 | -115.189 | 2000/5/31 | 2000/6/1 | 24.4 | Grass |
| MB | NW006 | 54.064 | -100.780 | 2015/5/25 | 2015/6/15 | 410 | Surface |
| ON | CHA16 | 48.376 | -82.843 | 2005/7/15 | 2005/7/17 | 0.7 | Crown |
| NT | CR-016 | 62.191 | -110.998 | 2004/7/12 | 2004/9/27 | 6977.53 | Forest |

In this project, the wildfire records in Alberta, Canada, will be utilised to explore and build forest fire prediction models. Alberta is the sixth-largest province in Canada, covering 662,000 square kilometres. It is also the province with third-most wildfires in Canadian history. The boreal forest covers the northern half of Alberta, and the Rocky Mountains on province's southern border are also mostly covered with forest. Conifers, rich in flammable branch lipids and oils, are the most common tree species in Alberta. After screening out the forest fire data from Alberta, Tableau was used to visualize the density, latitude histogram, and longitude histogram of historical forest fires in Alberta, as shown in Figure 2.

This figure shows the locations of wildfires happened in Alberta from 1950 to 2019 and the red areas have higher forest fire points density. As can be seen from density and histograms of

historical ignition points, most of the historical forest fires in Alberta occurred in central Alberta and the Rocky Mountains region in the southwest.

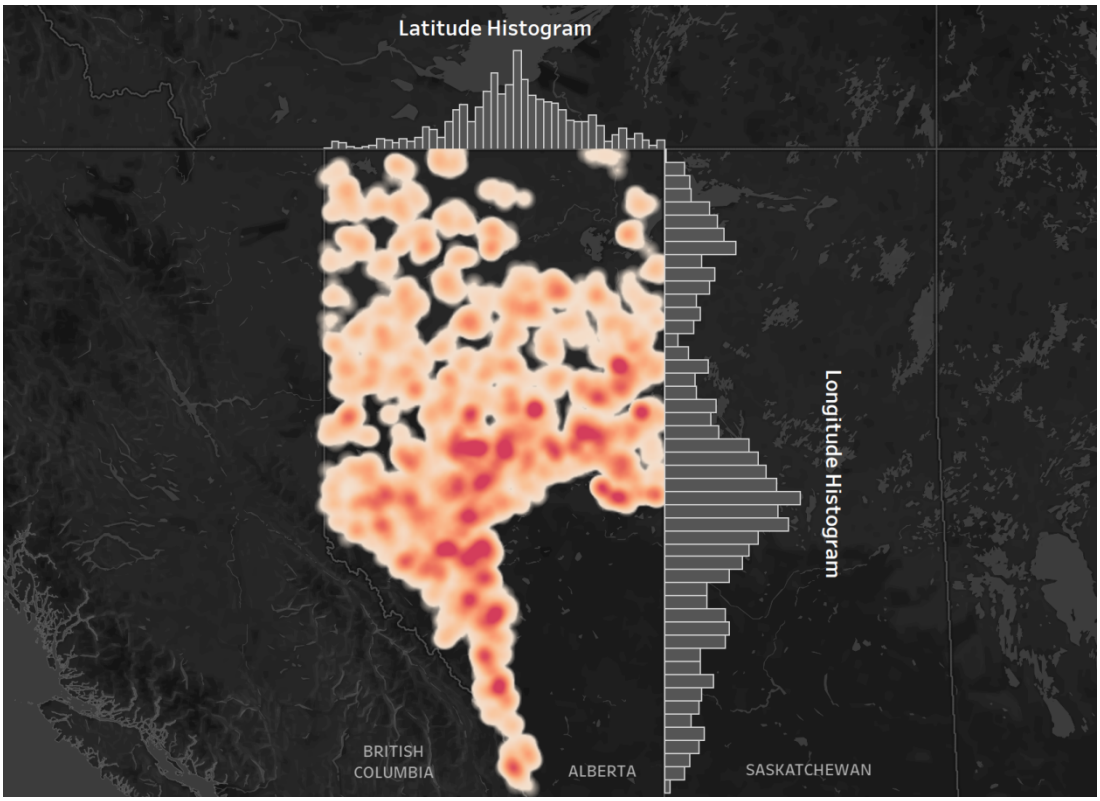


**Figure 2.** Historical Ignition Points Map

Pre-processing the historical forest fire dataset is the next stage. The original dataset contains a large number of null values due to missing or unmeasured data. This part of data will be deleted. Next, forest fire sizes will be ranked. According to the wildfire size classification system of America National Wildfire Coordinating Group, the size of forest fires can be classified into seven classes, as shown in Table 2.

**Table 2.** Forest fire classification

| Fire Size | Fire Area (acre) |
|---|---|
| Class A | ⩽0.25 |
| Class B | > 0.25 & ⩽10 |
| Class C | > 10 & ⩽100 |
| Class D | > 100 & ⩽300 |
| Class E | > 300 & ⩽1000 |
| Class F | > 1000 & ⩽5000 |
| Class G | > 5000 |

As a result, this project converted the original data's forest fire hectares to acres and ranked each record's fire size. The statistics of historical forest fires in different sizes are shown in Figure 3.

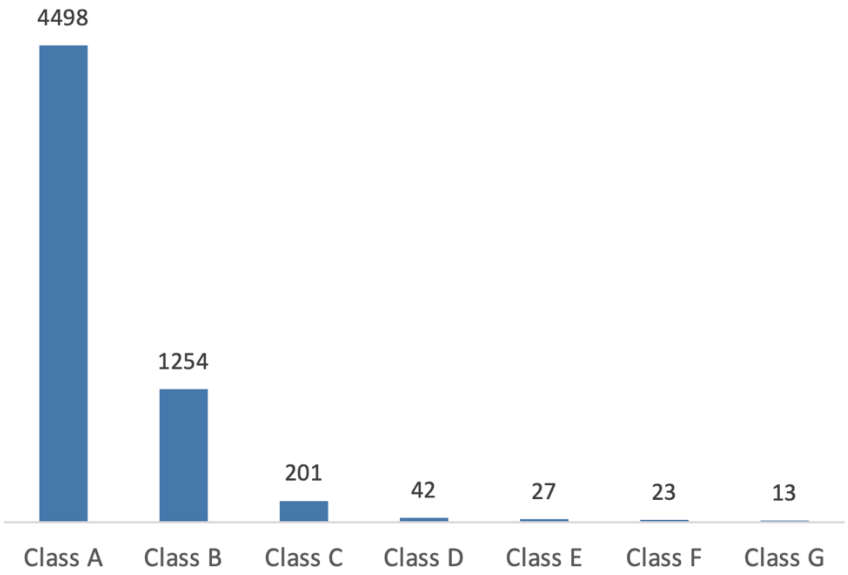


**Figure 3.** Statistics of historical ignition points in each fire size 1983-2019

As seen in the figure, the majority of wildfires in Alberta were Class A. Due to the small number of historical ignition points larger than Class D, the fire size is ranked to Class A, Class B, Class C, and ⩾Class D in this project.

6.1.2. The Weather Data of Historical Ignition Points

Historical weather dataset for this project was obtained from the Canadian Government's Past Weather and Climate website. This website contains past hourly, daily, and monthly weather records from 6,970 weather stations in Canada from 1990 to the present.

The attack date, longitude, and latitude in forest fire dataset are utilised to compile weather records. Due to the large volume of ignition data, this project uses the website crawling technique to collect meteorological dataset from weather stations nearest to the location of the historical forest fire point. In light of web crawler ethics, this project pre-investigated the Canada Climate website's robots policy and confirmed that web crawlers are permitted in the

Past Weather and Climate website. After collecting all the weather data of historical forest fires, the data with null values would be deleted. Table 3 shows samples of pre-processed ignition point climate data. The indexes are explained in Table 4. According to the literature review, key factors such as temperature, total precip, the direction of wind, and speed of wind will be chosen for feature selection and modelling.

**Table 3.** Samples of historical ignition point weather data

| FIRE_ID | MaxT | MinT | MeanT | HDD | CDD | TR | TS | TP | SG | DMG | SMG |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PWF131 | 7 | 5 | 26.6 | 12 | 19.3 | 0 | 1.3 | 1.2 | 0 | 1.2 | 0 |
| FS031 | 6 | 5 | 24.8 | 8.5 | 16.7 | 1.3 | 0 | 0 | 0 | 0 | 0 |
| MWF101 | 7 | 15 | 21.7 | 10.8 | 16.3 | 1.7 | 0 | 10.8 | 0 | 10.8 | 0 |
| LWF145 | 10 | 28 | 4 | -3.1 | 0.5 | 17.5 | 0 | 0 | 0 | 0 | 3 |
| G10386 | 3 | 15 | -1.9 | -8.2 | -5.1 | 23.1 | 0 | 0 | 1.2 | 0.8 | 25 |

**Table 4.** Index explanation

| Index | Explanation | Index | Explanation |
|---|---|---|---|
| MaxT | Maximum temperature (°C) | TS | Total snow (cm) |
| MinT | Minimum temperature (°C) | TP | Total precip (mm) |
| MeanT | Mean temperature (°C) | SG | Snow on ground (cm) |
| HDD | Heat deg days | DMG | Direction of maximum gust |
| CDD | Cool deg days | SMG | Speed of maximum gust (km/h) |

6.1.3. The Fuel Moisture Data of Historical Ignition Points

The fuel moisture dataset of historical ignition points is collected from Canada Wildland Fire Information System Datamart (2020). This dataset was recorded by province forest workers in collaboration with Natural Resources Canada and Environment and Climate Change Canada for fuel moisture related data on historical forest fires in Alberta. The fuel moisture dataset includes fire ID, relative humidity (RH), fine fuel moisture code (FFMC), duff moisture code (DMC), drought code (DC) and other dimensions of historical forest fires, as shown in Table 5.

**Table 5.** The fuel moisture data samples

| FIRE_ID | RH | FFMC | DMC | DC |
|---|---|---|---|---|
| SWF137 | 50.925 | 86.113 | 41.07 | 330.10 |
| G90343 | 55.261 | 86.590 | 12.04 | 216.12 |
| GWF035 | 41.810 | 88.908 | 33.14 | 316.19 |
| LWF090 | 23.700 | 93.900 | 57.40 | 531.90 |
| HWF058 | 22.412 | 93.960 | 60.14 | 201.68 |

Both fuel moisture dataset and historical fire dataset have the same fire ID. Therefore, this project connects fuel moisture dataset and historical fire dataset through $INNER\ JOIN$ keyword in MYSQL.

6.1.4. The Elevation Data of Historical Ignition Points

Because different elevations influence oxygen concentrations and related humidity, the change in elevation of forest fires has a crucial impact to the spread of wildfires. The Google Elevation API suggests an easy method to collect altitude data from numerous locations on the globe. All places on the earth's surface, even negative-valued deep sites on the sea floor, are available via the Google Elevation API. This project built a series of locations according to the longitude and latitude of historical forest fires, and then used the Google Elevation API to retrieve the altitude of each location.

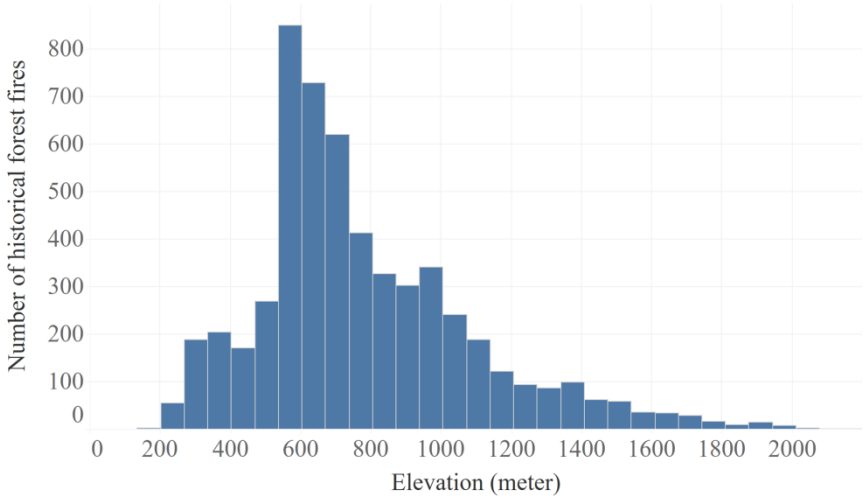


**Figure 4.** Historical ignition point elevation histogram

As shown in Figure 4, the occurrence of historical forest fires increased firstly and then decreased as elevation increased. The eastern plains of Alberta range from 500 to 700 meters, while the western Rocky Mountains range from 1,000 to 4,000 meters. The fire elevation histogram shows that most forest fires occurred in the plains, and a smaller number came from the lower Rocky Mountains.

6.1.5. Data Merging

The final step of data pre-processing is to combine the above historical ignition points dataset, weather dataset, fuel moisture dataset, and elevation dataset into one dataset. Because each dataset has the same fire ID, all variables can be stored in one dataset using the $JOINS$ keyword in MYSQL. After merging the data, the target and feature data for feature selection and modelling are retained in the dataset. The target data is fire size, and the feature data are elevation, temperature, relative humidity, speed of maximum gust, direction of maximum gust, 24-hour precip, fine fuel moisture code, duff moisture code and drought code. The final sample dataset is shown in the Table 6. The Table 7 describes the meaning of each feature data in the final dataset.

**Table 6.** Final dataset sample

| SIZE | EL | TEMP | RH | DMG | SMG | PRECIP | FFMC | DMC | DC |
|---|---|---|---|---|---|---|---|---|---|
| ≥Class D | 779 | 18.345 | 50.925 | 18 | 50 | 0.059 | 86.113 | 41.07 | 330.10 |
| Class A | 360 | 22.795 | 55.261 | 5 | 35 | 0.000 | 86.590 | 12.04 | 216.12 |
| Class B | 925 | 23.796 | 41.810 | 17 | 37 | 0.000 | 88.908 | 33.14 | 316.19 |
| ≥Class D | 662 | 28.800 | 23.700 | 18 | 45 | 0.000 | 93.900 | 57.40 | 531.90 |
| Class C | 881 | 27.482 | 22.412 | 8 | 39 | 0.000 | 93.960 | 60.14 | 201.68 |

**Table 7.** Feature explanation

| Feature | Description | Feature | Description |
|---|---|---|---|
| SIZE | The size of historical forest fires | DMG | The direction of maximum gust |
| EL | The elevation of historical ignition points | PRECIP | The 24-hour precip of ignition points |
| TEMP | The average temperature of ignition points | FFMC | The fine fuel moisture code |
| RH | The relative humidity of ignition points | DMC | The duff moisture code |
| SMG | The speed of maximum gust (km/h) | DC | The drought code |

### *6.2. Feature Selection*

Following data preparation, feature selection is the next step, and it has a significant influence on the model's performance. Feature selection is a technique for reducing the number of variables in a model by selecting features that are most beneficial in predicting the objective. Increasing the quantity of features within the model would enhance its predictive capacity, but only up to a point. The model's predictive performance would improve as the number of features increased, but after the feature number has passed its peak, the predictive performance will decline.

In this project, the $MinMaxScaler$ function in Python was used to data standardisation before feature selection. Because the features are on drastically different scales, if the data is not normalised, the efficiency of gradient descent during the training phase will reduce, and the accuracy of model training will also be affected. Therefore, the features were normalised to be at the same scale during training. The data normalisation formula is as follows. In this formula, $\max(X_i)$ is the maximum feature value, $\min(X_i)$ is the minimum feature value, and $X_i$ is the original feature value.

$$z = \frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)}$$

The $SelectFromModel$ function in Python was used to select the key features. $SelectFromModel$ is a meta-transformer that can be used with any estimator that can measure importance of each feature through a particular attribute, such as $coef\_$ and $feature\_importances\_$. If the importance of the feature is less than the $threshold$ parameter, the feature will be considered insignificant and eliminated. The Python functions used in feature selection step is show in Table 8.

**Table 8.** Function Description

| Function | Library | Description |
|---|---|---|
| MinMaxScaler | sklearn.preprocessing | Scale each feature to a given range |
| SelectFromModel | sklearn.feature_selection | Select features based on importance weights |
| coef_ | sklearn.feature_selection | Estimated coefficients for the algorithm |
| feature_importances_ | sklearn.feature_selection | Feature importances |
| get_support | sklearn.feature_selection | Get a integer index of the features selected |
| RandomForestClassifier | sklearn.ensemble | Random Forest meta estimator |
| SVC | sklearn.svm | Support vector classifier |
| LogisticRegression | sklearn.linear_model | Logistic Regression classifier |

6.2.1. Tree-based Algorithms Feature Selection

The feature selection of tree-based algorithms can effectively improve the prediction accuracy and running speed of Decision Tree and Random Forest algorithm. The $SelectFromModel$ and $RandomForestClassifier$ functions in Python were used to calculate the $feature\_importances\_$ of each feature. The importance value of each feature is shown in the figure 5.

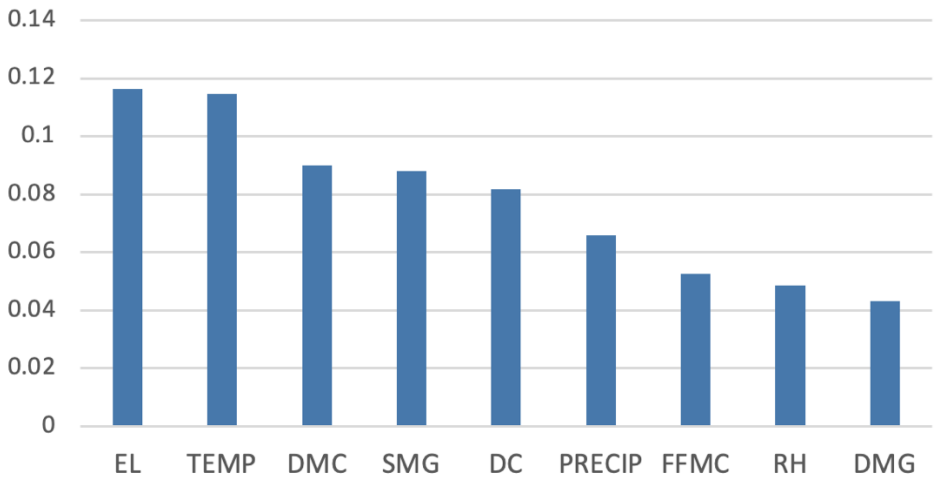


**Figure 5.** Feature importance of predictive model based on tree-based algorithm

The $feature\_importances\_$ of elevation, temperature, duff moisture code, speed of maximum gust and drought code are relatively high, exceeding 0.08. Therefore, these five

features will be used to train the Decision Tree and Random Forest-based wildfire size prediction models.

6.2.2. Logistic Regression Feature Selection

The $SelectFromModel$ and $LogisticRegression$ functions in Python are used to select features for the Logistic Regression algorithm. The $ceof\_$ of each feature is shown in the Figure 6.

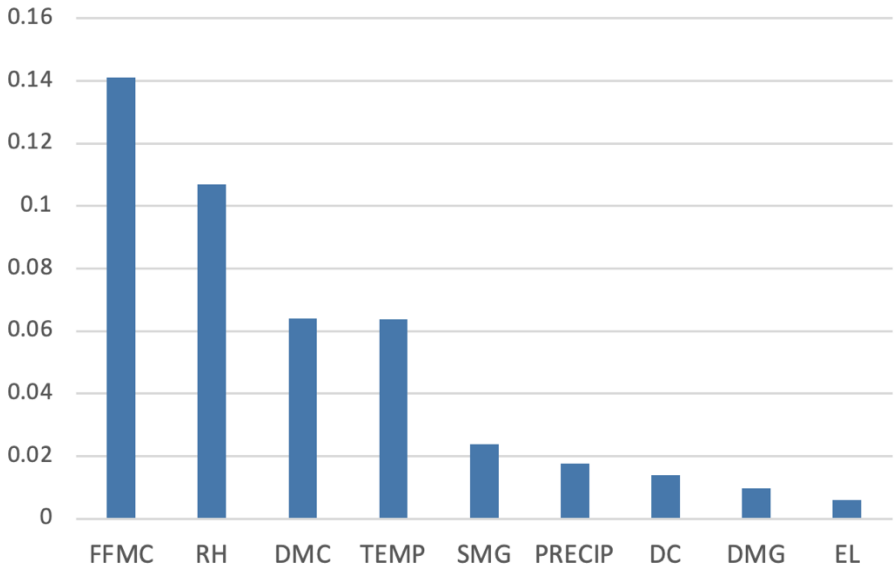


**Figure 6.** Feature importance of predictive model based on Logistic Regression

The $ceof\_$ of fine fuel moisture code, relative humidity, duff moisture code, and temperature are significantly higher than the others, indicating that these four features can effectively improve the running speed and prediction performance of the Logistic Regression algorithm model. Therefore, these four features will be employed to establish the forest fire prediction model based on the Logistic Regression algorithm.

6.2.3. Support Vector Machine Algorithm Feature Selection

The $SelectFromModel$ and $LinearSVC$ functions in Python are used to select features for the Support Vector Machine algorithm. The $ceof\_$ of each feature is shown in the Figure 7.

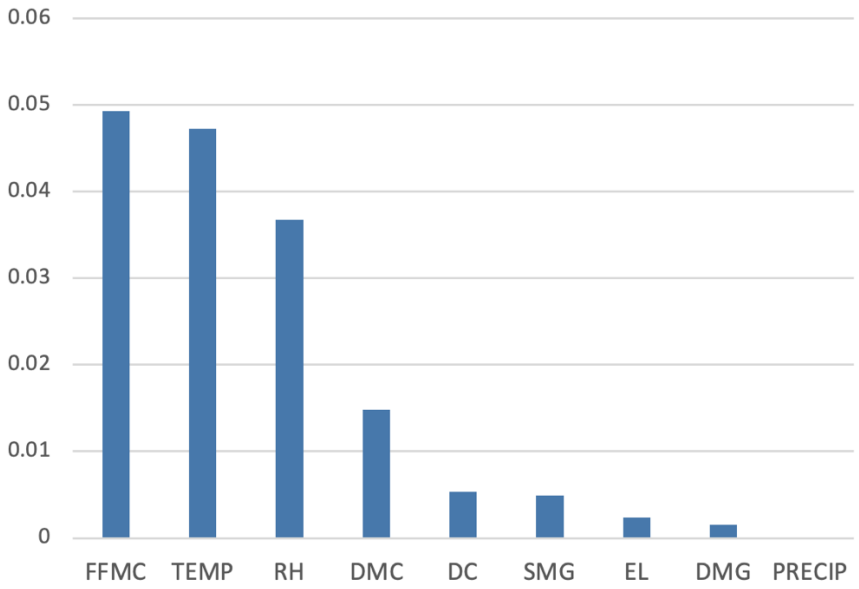


**Figure 7.** Feature importance of predictive model based on SVM

The fine fuel moisture code, temperature, relative humidity, and duff moisture code all have a higher $ceof\_$ than the others. It indicates that these four features can effectively increase the Support Vector Machine algorithm model's running speed and prediction performance. As a result, the forest fire prediction model based on the Support Vector Machine algorithm will be trained using these four features.

### *6.3. Modelling and Prediction*

The first step in modelling is to split the normalised dataset into training dataset and testing dataset. Taking tree-based algorithm modelling as an example, this project marks the features as X and the size as Y. The dataset for tree-based algorithm modelling is shown in Table 9.

**Table 9**. Normalised data samples for tree-based algorithms

| X | | | | | Y |
|---|---|---|---|---|---|
| **EL** | **TEMP** | **SMG** | **DMC** | **DC** | **FIRE_SIZE** |
| 0.361 | 0.981 | 0.432 | 0.267 | 0.510 | ≥Class D |
| 0.298 | 0.670 | 0.086 | 0.261 | 0.905 | Class A |
| 0.463 | 0.425 | 0.186 | 0.295 | 0.395 | Class B |
| 0.107 | 0.757 | 0.368 | 0.299 | 0.669 | ≥Class D |
| 0.125 | 0.898 | 0.196 | 0.367 | 0.554 | Class C |

The Scikit-learn library's $train\_test\_split()$ function is used to split datasets in this project, then 25% random data in the whole dataset is being used as testing dataset, while the rest is being used as training dataset. The training dataset's features (X_train) and fire size (Y_train) will be used to model training. After the model is trained, it will predict the fire size (Y_predict) based on the features (X_test) of the testing dataset. Finally, the projected fire sizes (Y_predict) will be compared to the actual fire sizes (Y_test) in the testing dataset to evaluate the model's prediction performance. The functions used in modelling is shown in Table 10.

**Table 10.** Function description

| Function | Library | Description |
|---|---|---|
| train_test_split | sklearn.model_selection | Split data into random train and test subsets |
| GridSearchCV | sklearn.model_selection | Exhaustive search over specified parameter |
| DecisionTreeClassifier | sklearn.tree | Decision Tree classifier. |
| RandomForestClassifier | sklearn.ensemble | Random Forest meta estimator |
| LogisticRegression | sklearn.linear_model | Logistic Regression classifier |
| SVC | sklearn.svm | Support vector classifier |

### *6.4. Predicted Results Analysis*

#### 6.4.1. Evaluation Method

The Accuracy, Precision, Recall and F1-score are employed to evaluate the prediction performance of models in this project.

The percentage of correct predictions in the prediction results, known as Accuracy, is commonly used to indicate the model's predictive ability. Accuracy, however, can be misleading when the data samples are unbalanced. For example, Class A forest fires in this dataset account for 74%. Based on this dataset, if a model predicted all fires as Class A, it would have 74% accuracy, but in fact, the predictive ability of this model is terrible.

Precision is a measure of the relevance of the results, whereas Recall is a measure of the ratio of correctly relevant results returned. The F1-score is a balanced reflection of model predictive performance, as it is the weighted mean of Precision and Recall. Therefore, Precision, Recall

and F1-score can effectively evaluate the predictive performance of models with imbalanced samples. The indexes used in model evaluation are shown in Table 11.

**Table 11.** Evaluation indexes description

| Index | Description |
|---|---|
| Accuracy | Accuracy is the percentage of correctly predicted observations to the overall dataset. Accuracy ranges from 0 to 1. Accuracy closer to 1 indicates better predictive ability. |
| Precision | Precision is the percentage of correctly predicted positive observations to all predicted positive observations. Precision ranges from 0 to 1. Precision closer to 1 indicates better predictive ability. |
| Recall | Recall is the percentage of correctly predicted positive observations to all observations in actual class. Recall ranges from 0 to 1. Recall closer to 1 indicates better predictive ability. |
| F1-score | The F1-score is the weighted average of Precision and Recall. The F1-score indicates the balance between precision and Recall. F1-score ranges from 0 to 1. F1-score closer to 1 indicates better predictive ability. $F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$ |

6.4.2. Predictive Performance Evaluation

The prediction performance evaluation of Decision Tree, Random Forest, Support Vector Machine and Logistic Regression model is shown in the Figure 8.

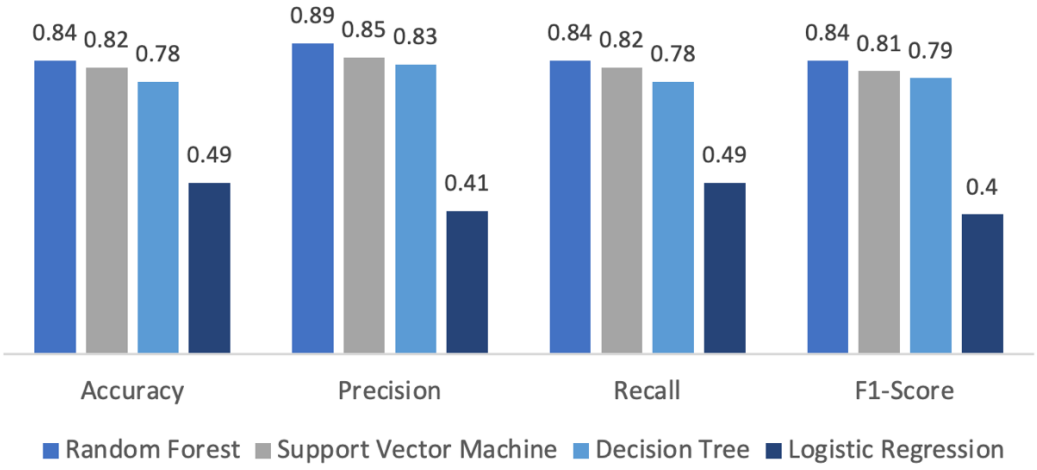


**Figure 8.** Bar chart of predictive performance evaluation

As can be seen from the figure, the predictive performance of Logistic Regression is significantly lower than that of the other three models. Decision Tree, Random Forest, and

Support Vector Machine model all have about 80% predictive performance evaluation. Among them, the Random Forest model came out on top in all the indexes, followed by the Support Vector Machine model and Decision Tree model.

To further study the prediction performance of models in certain forest fires size, Table 12 shows the precision, recall and F1-score of models in each fire-size.

**Table 12.** Predictive performance evaluation on each fire-size

| Class | Model | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Class A | DT | 0.67 | 0.67 | 0.67 |
| | RF | 1.00 | 0.50 | 0.67 |
| | LR | 0.14 | 0.33 | 0.20 |
| | SVM | 1.00 | 0.33 | 0.50 |
| Class B | DT | 0.90 | 0.75 | 0.82 |
| | RF | 0.91 | 0.88 | 0.89 |
| | LR | 0.67 | 0.33 | 0.44 |
| | SVM | 0.79 | 0.96 | 0.87 |
| Class C | DT | 0.79 | 0.96 | 0.87 |
| | RF | 1.00 | 0.80 | 0.89 |
| | LR | 0.62 | 0.53 | 0.57 |
| | SVM | 1.00 | 0.80 | 0.89 |
| ≥ Class D | DT | 0.53 | 0.90 | 0.67 |
| | RF | 0.59 | 1.00 | 0.74 |
| | LR | 0.12 | 0.20 | 0.15 |
| | SVM | 0.67 | 0.80 | 0.73 |

In Class A, both Decision Tree and Random Forest model have relatively high prediction performance evaluation, indicating that the tree-based models have better prediction ability for small forest fire. In Class B, the Random Forest model has the higher F1-score, reaching 0.89, followed by the Support Vector Machine model with 0.87. In Class C, the Support Vector Machine model and Random Forest model have almost identical predicted performance with 0.89 F1-score. In ≥ Class D, the Random Forest model has the higher F1-score, reaching 0.74, followed by the Support Vector Machine model with 0.73. Furthermore, the Logistic Regression model has poor predictive performance evaluation in all four classes.

In summary, both the tree-based and Support Vector Machine models can effectively predict the fire size based on the related forest fire feature. And the Random Forest algorithm-based

predictive model has the highest predictive evaluation. Logistic Regression performed poorly in predicting forest fire scale, indicating that the relationship between forest fire size and features was not linear.

## 7. Ethical Implications and Limitations

### *7.1. Ethical Implications*

When collecting data for this project, data crawling ethics were carefully considered. This project uses web crawler technologies to collect historical climate data of forest fire points from Alberta. The robots exclusion standard is the main morality norm in global Internet community. This standard is a web crawler and other online robot communication protocol for webpages. This standard describes which parts of a website should not be scanned or downloaded by a web crawler.

Therefore, before crawl the historical weather data, the robots exclusion standard of Canadian Government's Past Weather and Climate website was investigated in this project. This website's robots exclusion standard is shown below:

$$User - agent: *$$

$$Disallow:$$

The * after $User - agent$ means all robots and web crawlers are allowed on the website. The $Disallow$ with no value indicates that all pages on the website are available. Therefore, the Past Weather and Climate website's standard means that all robots and web crawlers are allowed to read all pages on this website. Furthermore, the meteorological data gathered from the website will be utilised solely for personal study, and data containing sensitive information will not be investigated in this project. Therefore, the data collection methods and data sources involved in this project comply with the standard and ethics of the Internet

### *7.2. Limitations*

This project compiles the historical weather data from 6970 weather stations across Canada. According to the longitude and latitude of the historical wildfire points, this project crawls historical weather data from the nearest weather station. As a result, there is a discrepancy

between the weather data obtained through this method and the actual weather data at fire points.

This project mainly considers weather, elevation, and fuel moisture as the feature of the prediction model. Still, some other factors also affect the size of forest fires, such as population density, vegetation type, and slope aspect. Because some of these data contain information about a country's geography and population, they are not available due to security or ethical considerations. If all these factors could be studied or incorporated into the modelling, this project's thoroughness and prediction performance of the model would be improved further.

Many other algorithms are also worth applying to the wildfire prediction research. For example, the geographically weighted regression algorithm can avoid the discrepancy caused by spatial heterogeneity. In addition, neural networks and other algorithms in deep learning can also predict the size of forest fires. These algorithms have their own advantages and merit further research.

## 8. Conclusions and Outlook

### *8.1. Conclusions and Recommendations*

With global warming, forest fires are increasingly harmful to the environment, property, and human health. Early Forecasting of wildfires' eventual magnitude could assist fire departments in formulating effective rescue plans to reduce losses. Therefore, the main objective of this project is to develop a more accurate model that uses fire-related factors as input and forest fire size as output to predict the size of forest fires. The experimental datasets used in this project include the historical ignition point dataset of Alberta from 1950 to 2019 downloaded from Canada's Natural Resources Datamart, the historical ignition point humidity dataset collected from Canada Wildland Fire Information System Datamart, the historical ignition point weather dataset crawled from the Canadian Weather website, and the ignition point elevation dataset collected from the Google Elevation API. The following are the main research findings of this project.

(1) Through the study of the world-famous Fire Danger Rating Systems and the literature review of forest fires related factors, this project found that the main factors affecting forest fire are temperature, wind speed, relative humidity, precip, fine fuel moisture code, duff moisture code and drought code. Through the overview of forest fire prediction methods, this project found that Logistic Regression and tree-based algorithms are two widely used and effective algorithms in wildfire forecast research. In addition, since the forest fire size prediction is a classification issue, some classification algorithms in machine learning, like the Support Vector Machine algorithm, could also be used to research the fire size prediction.

(2) Based on historical fire data from Alberta, this project found that forest fires occurred mainly in Alberta's central plains and southern Rocky Mountain regions at altitudes of 600

to 1200 meter. In Alberta, historical forest fire sizes are dominated by Class A, accounting for 74% of all forest fires. In the feature selection research of different algorithms, this project found that the key features of tree-based algorithms are elevation, temperature, duff moisture code, speed of maximum gust and drought code. The key features of the Logistic Regression algorithm are fine fuel moisture code, relative humidity, duff moisture code, and temperature. The key features of the Support Vector Machine algorithm are fine fuel moisture code, temperature, relative humidity, and duff moisture code.

(3) Through the evaluation of model prediction performance, the project found that the model based on Random Forest algorithm has a superior predictive performance in general, with an F1-score of 0.84, followed by the Support Vector Machine model with an F1-score of 0.81. Nevertheless, the predictive performance of the Logistic Regression model is lower, indicating the relationship between related features and the magnitude of forest fires is not linear. Therefore, it can be concluded that both the Random Forest algorithm and Support Vector Machine algorithm can predict the final fire size based on fire-related features with reasonable accuracy.

In light of the increasing severity of forest fire dangers, this project recommends that forest fire departments and researchers to develop forest fire prediction models using the Random Forest algorithm based on key fire-related features, including elevation, temperature, duff moisture code, speed of maximum gust and drought code. When a forest fire occurs, these key fire-related factors can be immediately measured by the local weather bureau or forest service. Before modelling, it is recommended to select key features by comparing the correlation coefficients between features and objective. This feature selection method improves the efficiency and accuracy of the model by reducing the quantity of features. In the model performance evaluation step, Precision, Recall and F1-Score are recommended to evaluate the prediction ability of the model if the samples are unbalanced. In this project's experiment, the

Random Forest model has a better predictive performance than other algorithms for historical forest fire size in Alberta. Therefore, this method will effectively help fire departments to understand the size of a fire and formulate effective rescue measures at the early stage of the forest fire.

### *8.2. Outlook*

Human activities are causing an increasing number of forest fires (Flannigan et al., 2001). Therefore, it is necessary to find an appropriate method for quantifying the impact of human activities to the forest fires' occurrence in future studies. For example, social and economic variables, such as population density, the proximity of roads and rivers, and the distribution of electric wires, can be added to the prediction model as a quantitative representation of human activities. However, due to the complexity of human activities, it is difficult to describe the above indicators accurately. In addition, storing quantitative data of human activities into the database is also a big challenge for current technology. Analysing the quantitative effect of human activities on forest fires needs further research.

The prediction of forest fire using a hybrid artificial intelligence model is also worth further study. A single prediction method may lack robustness because small changes in the model or data can lead to large differences in prediction results. Hybrid artificial intelligence algorithm models based on artificial intelligence methods and optimization algorithms are widely used in commercial and financial fields but rarely used in forest fire prediction modelling. Therefore, introducing the hybrid model of artificial intelligence algorithm and optimization algorithm into the forest fire size forecast model to increase performance and robustness of prediction ought to be explored in future studies.

**References**


[1] Weisse, M., & Goldman, E. (2021). Primary rainforest destruction increased 12% from 2019 to 2020. Forest Pulse: World Resources Institute. https://research. wri. org/gfr/forest-pulse.

[2] Forest Research (Ed.) (2021). Woodland Fires. https://www.forestresearch.gov.uk/tools-and-resources/statistics/forestry-statistics/forestry-statistics-2018/environment/woodland-fires/

[3] US Geological Survey (2006). Wildfire Hazards—A National Threat. https://pubs.usgs.gov/fs/2006/3015/2006-3015.pdf

[4] Insurance Information Institute (Ed.) (2021). Facts + Statistics: Wildfires. https://www.iii.org/fact-statistic/facts-statistics-wildfires

[5] Climate Reality Project (Ed.) (2020). GLOBAL WILDFIRES BY THE NUMBERS. https://www.climaterealityproject.org/blog/global-wildfires-numbers

[6] DuBois, Coert. (1914). Systematic Fire Protection in the California Forests. U. S. Department of Agriculture. Forest service.

[7] Wright, J. G. (1937). Preliminary improved fire hazard index tables for pine forests at Petawawa Forest Experiment Station. Department of Mines and Resources.

[8] Gisborne, H. T. (1936). Measuring fire weather and forest inflammability (No. 398). US Department of Agriculture.

[9] Deeming, J. E., Burgan, R. E., & Cohen, J. D. (1977). The national fire-danger rating system, 1978 (Vol. 39). Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station.

[10] Jolly, W. M., Graham, J. M., Michaelis, A., Nemani, R., & Running, S. W. (2005). A flexible, integrated system for generating meteorological surfaces derived from point

sources across multiple geographic scales. Environmental modelling & software, 20(7), 873-882.

[11] Nelson Jr, R. M. (2000). Prediction of diurnal change in 10-h fuel stick moisture content. Canadian Journal of Forest Research, 30(7), 1071-1087.

[12] Burgan, R. E. (1996). Using NDVI to assess departure from average greenness and its relation to fire business (Vol. 333). US Department of Agriculture, Forest Service, Intermountain Research Station.

[13] Van Wagner, C. E., & Forest, P. (1987). Development and structure of the canadian forest fireweather index system. In Can. For. Serv., Forestry Tech. Rep.

[14] McAlpine, R. S. (1990). Seasonal trends in the Drought Code component of the Canadian Forest Fire Weather Index System (No. PI-X-97E/F).

[15] McArthur, A. G. (1958, July). The preparation and use of fire danger tables. In Fire Weather Conference (pp. 15-17).

[16] San-Miguel-Ayanz, J., Carlson, J. D., Alexander, M., Tolhurst, K., Morgan, G., Sneeuwjagt, R., & Dudley, M. (2003). Current methods to assess fire danger potential. In Wildland fire danger estimation and mapping: The role of remote sensing data (pp. 21-61).

[17] Nakicenovic, N., Alcamo, J., Davis, G., Vries, B. D., Fenhann, J., Gaffin, S., ... & Zhou, D. (2000). Special report on emissions scenarios.

[18] Allen, C. D., Macalady, A. K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., ... & Cobb, N. (2010). A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. Forest ecology and management, 259(4), 660-684.

[19] Batllori, E., Parisien, M. A., Krawchuk, M. A., & Moritz, M. A. (2013). Climate change-induced shifts in fire for M editerranean ecosystems. Global Ecology and Biogeography, 22(10), 1118-1129.

[20] Wibbenmeyer, M., & McDarris, A. (2021). Wildfires in the United States 101: Context and Consequences.

[21] European Environment Agency (Ed.) (2021). Forest fires. European Environment Agency. https://www.eea.europa.eu/data-and-maps/indicators/forest-fire-danger-3/assessment

[22] Turco, M., Bedia, J., Di Liberto, F., Fiorucci, P., Von Hardenberg, J., Koutsias, N., ... & Provenzale, A. (2016). Decreasing fires in mediterranean Europe. PLoS one, 11(3), e0150663.

[23] National Forestry Database (Ed.) (2020). Forest area burned and number of forest fires. http://nfdp.ccfm.org/en/data/fires.php

[24] Natural Resources Canada (Ed.) (2020). Indicator: Forest fires. How does disturbance shape Canada's forests? https://www.nrcan.gc.ca/our-natural-resources/forests/state-canadas-forests-report/disturbance-canadas-forests/indicator-forest-fires/16392

[25] Sano, I., Mukai, S., Nakata, M., Holben, B. N., & Kikuchi, N. (2011, October). Optical properties of biomass burning aerosols during Russian forest fire events in 2010. In Remote Sensing of Clouds and the Atmosphere XVI (Vol. 8177, p. 81770D). International Society for Optics and Photonics.

[26] Catchpole, E. A., Catchpole, W. R., & Rothermel, R. C. (1993). Fire behavior experiments in mixed fuel complexes. International Journal of Wildland Fire, 3(1), 45-57.

[27] Chuvieco, E., Cocero, D., Riano, D., Martin, P., Martınez-Vega, J., De La Riva, J., & Pérez, F. (2004). Combining NDVI and surface temperature for the estimation of live

fuel moisture content in forest fire danger rating. Remote Sensing of Environment, 92(3), 322-331.

[28] Flannigan, M. D., Wotton, B. M., Marshall, G. A., De Groot, W. J., Johnston, J., Jurko, N., & Cantin, A. S. (2016). Fuel moisture sensitivity to temperature and precipitation: climate change implications. Climatic Change, 134(1), 59-71.

[29] Stocks, B. J., Lynham, T. J., Lawson, B. D., Alexander, M. E., Wagner, C. V., McAlpine, R. S., & Dube, D. E. (1989). Canadian forest fire danger rating system: an overview. The Forestry Chronicle, 65(4), 258-265.

[30] Amiro, B. D., Stocks, B. J., Alexander, M. E., Flannigan, M. D., & Wotton, B. M. (2001). Fire, climate change, carbon and fuel management in the Canadian boreal forest. International Journal of Wildland Fire, 10(4), 405-413.

[31] Lawson, B. D., Armitage, O. B., & Hoskins, W. D. (1996). Diurnal variation in the Fine Fuel Moisture Code: tables and computer source code. FRDA report.

[32] Wotton, B. M., & Martell, D. L. (2005). A lightning fire occurrence model for Ontario. Canadian Journal of Forest Research, 35(6), 1389-1401.

[33] Wotton, B. M., & Beverly, J. L. (2007). Stand-specific litter moisture content calibrations for the Canadian Fine Fuel Moisture Code. International Journal of Wildland Fire, 16(4), 463-472.

[34] Wagner, C. V. (1979). A laboratory study of weather effects on the drying rate of jack pine litter. Canadian Journal of Forest Research, 9(2), 267-275.

[35] Byram, G. M., & Jemison, G. M. (1943). Solar radiation and forest fuel moisture. Journal of Agricultural Research, 67(4), 149-176.

[36] Marsden-Smedley, J. B., Catchpole, W. R., & Pyrke, A. (2001). Fire modelling in Tasmanian buttongrass moorlands. IV. Sustaining versus non-sustaining fires. International Journal of Wildland Fire, 10(2), 255-262.

[37] Plucinski, M. P., Anderson, W. R., Bradstock, R. A., & Gill, A. M. (2010). The initiation of fire spread in shrubland fuels recreated in the laboratory. International Journal of Wildland Fire, 19(4), 512-520.

[38] Viegas, D. X., Viegas, M. T. S. P., & Ferreira, A. D. (1992). Moisture content of fine forest fuels and fire occurrence in central Portugal. International Journal of Wildland Fire, 2(2), 69-86.

[39] Renkin, R. A., & Despain, D. G. (1992). Fuel moisture, forest type, and lightning-caused fire in Yellowstone National Park. Canadian Journal of Forest Research, 22(1), 37-45.

[40] Nash, C. H., & Johnson, E. A. (1996). Synoptic climatology of lightning-caused forest fires in subalpine and boreal forests. Canadian Journal of Forest Research, 26(10), 1859-1874.

[41] Ray, D., Nepstad, D., & Moutinho, P. (2005). Micrometeorological and canopy controls of fire susceptibility in a forested Amazon landscape. Ecological Applications, 15(5), 1664-1678.

[42] Chuvieco, E., González, I., Verdú, F., Aguado, I., & Yebra, M. (2009). Prediction of fire occurrence from live fuel moisture content measurements in a Mediterranean ecosystem. International Journal of Wildland Fire, 18(4), 430-441.

[43] Wright, H. A., & Bailey, A. W. (1982). Fire ecology: United states and southern canada. John Wiley & Sons.

[44] Viegas, D. X., Piñol, J., Viegas, M. T., & Ogaya, R. (2001). Estimating live fine fuels moisture content using meteorologically based indices. International Journal of Wildland Fire, 10(2), 223-240.

[45] Fried, J. S., Torn, M. S., & Mills, E. (2004). The impact of climate change on wildfire severity: a regional forecast for northern California. Climatic change, 64(1), 169-191.

[46] Harrington, P. (2012). Machine learning in action. Simon and Schuster.

[47] Prasad, V. K., Badarinath, K. V. S., & Eaturu, A. (2008). Biophysical and anthropogenic controls of forest fires in the Deccan Plateau, India. Journal of environmental management, 86(1), 1-13.

[48] Vega-García, C., Tatay-Nieto, J., Blanco, R., & Chuvieco, E. (2010). Evaluation of the influence of local fuel homogeneity on fire hazard through Landsat-5 TM texture measures. Photogrammetric Engineering & Remote Sensing, 76(7), 853-864.

[49] Fotheringham, S., Charlton, M., & Brunsdon, C. (1996). The geography of parameter space: an investigation of spatial non-stationarity. International Journal of Geographical Information Science, 10(5), 605-627.

[50] Rodrigues, M., de la Riva, J., & Fotheringham, S. (2014). Modeling the spatial variation of the explanatory factors of human-caused wildfires in Spain using geographically weighted Logistic Regression. Applied Geography, 48, 52-63.

[51] Coffield, S. R., Graff, C. A., Chen, Y., Smyth, P., Foufoula-Georgiou, E., & Randerson, J. T. (2019). Machine learning to predict final fire size at the time of ignition. International Journal of Wildland Fire, 28(11), 861–873. https://doi-org.uoelibrary.idm.oclc.org/10.1071/WF19023

[52] Qin, L., Shao, W., Du, G., Mou, J., & Bi, R. (2021). Predictive Modeling of Wildfires in the United States. 2021 2nd International Conference on Computing and Data Science (CDS), Computing and Data Science (CDS), 2021 2nd International Conference on, CDS, 562–567. https://doi-org.uoelibrary.idm.oclc.org/10.1109/CDS52072.2021.00102

[53] Collins, L., McCarthy, G., Mellor, A., Newell, G., & Smith, L. (2020). Training data requirements for fire severity mapping using Landsat imagery and Random Forest. Remote Sensing of Environment, 245. https://doi-org.uoelibrary.idm.oclc.org/10.1016/j.rse.2020.111839

[54] Archibald, S., Roy, D. P., van Wilgen, B. W., & Scholes, R. J. (2009). What limits fire? An examination of drivers of burnt area in Southern Africa. Global Change Biology, 15(3), 613-630.

[55] Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A., & Pereira, J. M. (2012). Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. Forest Ecology and Management, 275, 117-129.

[56] Miranda, B. R., Sturtevant, B. R., Stewart, S. I., & Hammer, R. B. (2011). Spatial and temporal drivers of wildfire occurrence in the context of rural development in northern Wisconsin, USA. International Journal of Wildland Fire, 21(2), 141-154.

[57] McCoy, V. M., & Burn, C. R. (2005). Potential alteration by climate change of the forest-fire regime in the boreal forest of central Yukon Territory. Arctic, 276-285.

[58] Podur, J. J., Martell, D. L., & Stanford, D. (2010). A compound Poisson model for the annual area burned by forest fires in the province of Ontario. Environmetrics, 21(5), 457-469.

[59] Chang Y, Zhu Z L, Bu R C, et al. Predicting fire occurrence patterns with logistic regression in Heilongjiang Province, China. Landscape Ecology, 2013, 28(10): 1989~2004

[60] Flannigan, M., Campbell, I., Wotton, M., Carcaillet, C., Richard, P., & Bergeron, Y. (2001). Future fire in Canada's boreal forest: paleoecology results and general circulation model-regional climate model simulations. Canadian journal of forest research, 31(5), 854-864.

[61] National Wildfire Coordinating Group (Ed.) (2009). NWCG Data Standard Fire Size Class Code Standard Data Values. https://www.nwcg.gov/data-standards/approved/fire-size-class

[62] Natural Resources Canada (2020). Historical ignition point data source. https://cwfis.cfs.nrcan.gc.ca/datamart

[63] Canadian Wildland Fire Information System Datamart (2020). Alberta Smoke Plume Observations data. https://cwfis.cfs.nrcan.gc.ca/datamart/metadata/absmoke