National College of Ireland

Analysis of Wildfire Risk Using Machine Learning and Distributed Computing in Canadian Regions

MSc Research Project
Data Analytics

Emmanuel Amadi
Student ID: x18178103

School of Computing
National College of Ireland

Supervisor: Dr. Cristina Muntean

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Emmanuel Amadi
Student ID:	X18178103
Programme:	Data Analytics **Year:** 2019
Module:	MSc Research Project
Supervisor:	Dr. Cristina Muntean
Submission Due Date:	12/12/2019
Project Title:	Analysis of wildfire risk using machine learning and distributed computing in Canadian regions
Word Count:	6654 **Page Count** 23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: .....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Analysis of Wildfire Risk Using Machine Learning and Distributed Computing in Canadian Regions

Emmanuel Amadi
X18178103

Abstract

Wildfires present a great danger to human lives and their environments. Early detection and rapid spread of a wildfire is a major challenge to some countries, especially during the summer period which must be reduced to prevent economic, ecological and social damage to human lives. Data mining algorithms can be applied to historic and near real-time data to gain useful insight that will aid the fire managers in predicting, reducing the cost of moving water tankers with heavy fire equipment and the tendency of the fire to spread if not quenched on time. The aim of this research is to investigate using unsupervised and supervised machine learning algorithms built on distributed computing in predicting and staging firefighting assets as close to where wildfires are likely to occur based on wildfire dataset. The method followed a knowledge discovery and data mining approach extracting insight from the NASA wildfire dataset to predict the occurrence of wildfire and reduce computational time. Consequently, this research methodology was implemented to achieve this by building wildfire models from remote sensing satellite data acquired from the Moderate Resolution Imaging Spectroradiometer (MODIS). Experimental results showed that K-means clustering with a silhouette score of 65% and random forest with reduced RMSE of 0.13 when treated as a regression analysis while for classification, the model gave high prediction accuracy of 97% and training time of 7 seconds. The results and performance of these models were determined using cross-validation, root mean square error (RMSE), R-squared and classification metrics.

Keywords: Wildfire prediction, Machine learning, Random forest, Classification algorithms, Kmeans clustering, Distributed computing, Regression algorithms.

1 Introduction

Wildfire over the years has caused significant environmental problems with economic, social and ecological adverse effects. Avoiding wildfire as early as possible is crucial to reduce the risk of human lives and properties as these wildfires are often caused by human negligence to the environment while other factors like lightning and topography also contribute to the occurrence and spread. Researches are constantly identifying the importance including more data points like metrological data points to predict the occurrence and plan strategies to mitigate the risk to human life. The complicated topography, a combination of complex fuel structure in the forest and varying metrological situation presents monitoring challenges in modelling and prediction of wildfire.

In Canada, a survey reveals that over 2.7 million hectares of land have been lost to 5,760 forest fires for the past 10 years. Early detection is vital to avoid fire by modelling a preparatory planning strategy using data mining techniques (Divya *et al.*, 2014). The Canadian Forest weather index used for simple calculation and modelling was transposed from metrological data into numerical indices for fire risk reduction. Improving the forest weather index is a vital resource for land agencies resource for fire scientists and land agencies globally, these improvements depend on a proper understanding of other scientific, weather and human factors influencing the occurrence and spread of wildfire (Collins *et al*, 2018). The recent trend in data mining and distributed computing is utilized to reduce the number of resources allocated to manage wildfire and address the computational time as well as monitoring challenges of collecting data over an active fire to aid in decision making to save the lives of the ground crew and the residents in the area. The key challenge in predicting wildfire is improving the predictive accuracy of the fire models, prompting more weather parameters to be included in the computation stressing the importance of big data for computation and analyzing this data comes with increased computational time (Cortez and Morais,2007). The fire managers often face the challenge of moving heavy equipment and water tankers anytime an alarm of wildfire is raised but the rapid spread of a wildfire is sometimes difficult to manage as a collection of real-time data is difficult due to inaccessible and dangerous environments. This implies that firefighting assets should be placed strategically by analyzing historical data to rapidly respond to fire alarm and also reduce prediction error as well as reduce the cost of moving this heavy equipment and reducing the threat to the lives of the firefighters and residents in the area.

Addressing the challenges stated above, this research aims to analyze fire locations in a bid to identify strategic locations to deploy firefighting equipment and cub the issues of fire exploding in size if not suppressed on time while considering the likelihood of fire to occur in these locations. Therefore, the research question ***To what extent can K-means clustering, regression, random forest and navie bayes built on distributed computing effectively predict wildfire in Canadian regions?*** is posed with the following objectives

- To improve the computational time and accuracy of wildfire predictive models
- Effectively process remote sensing data using a machine learning algorithm
- Validate the accuracy of different machine learning algorithms to find the best performing model.

Specifically, the contributions of this research are:

- The novelty of introducing k-mean for individual natural clusters to analyze the clusters based on historic data.
- Suggested a new computational method to analyze the data based on a distributed computing environment rather than the traditional method previously deployed.
- A new approach to a wildfire risk analysis based on distributed computing and machine learning rather than the traditional approach of using data mining algorithms alone previously deployed by researchers in this domain.

The remainder of this research is organized as follows: Section 2 reviews the relevant literature of our research, section 3 describes the research methodology, section 4 highlights

the design implementation for the techniques, analytical methods employed for this research, section 5 presents the experimental results conducted for this research, section 6 covers the evaluation and discussion of the results in detail and the research is then concluded in section 7.

2 Related Work

This research was built on existing ideas from recent, relevant and related papers by evaluating different methodologies previously adopted that significantly add to the wildfire research area and data analytics domain. The literature review section for this research compares different traditional and data mining techniques been developed in wildfire research with an overview of objectives and the limitations in achieving their results. This section is divided as follows: the first subsection provides insight to detecting wildfire using wireless sensor network and the relevant indicators that played important roles in achieving the results, the second section describes the use of machine learning techniques in predicting and managing wildfire risk by researchers using various indicators, the third section provides an overview of the study area considered with type of input data used and the last section concludes by evaluating and summarizing the key findings to further support the novelty and justify the need of this research.

2.1 Wildfire Detection Using Wireless Sensor Networks

In the prediction of wildfire, wireless sensor networks technologies are used to detect fire by setting fire thresholds and defining specific parameters like temperature and relative humidity is predefined. Wildfire that are managed with multiple resources are mainly controlled as the hazard increases with suppression tactics (Young *et al.*, 2019).This tactics are set as threshold which triggers an alarm when the reading of the threshold value is not the same as that of the sensors (Kansal *et al.*, 2016). This subsection describes the use of a wireless sensor network in wildfire risk management and data points consider necessary for this research.

Wireless Sensor Networks (WSN) consists of sensors used to monitor environmental or physical conditions. This technology can be applied to detect or predict wildfire in real-time. Singh *et al*., (2013) applied WSN placed on distributed locations and transmit information to the base station where an ensemble distributed machine learning approach is applied to the data for decision making. The approach uses two phases to detect an event in a distributed manner, the base and metaphase and the signals are clustered with SVM used at the base station detect and predict wildfire. The model improved energy efficiency in a distributed environment while improving the performance and time but the type of data points collected by the WSN was not specific to enable improvement on the model. Kansal *et al*., (2016) introduced the regression model approach to wildfire detection by dividing the dataset and performing analysis on them. The approach achieved high R-square value and low root mean square error on a small sample of forest fire data. The chances of the divided data to contain outliers are high which might cause the true representation of the data to be lost in cases with a large amount of data. Serna, Bermudez, and Casado, (2013) research was focused on developing a model that represents the entire fire area in circles and distributes each fire front

representation to firefighters in the area. The information is distributed wirelessly in the network and each firefighter maintains a fire representation that corresponds to a list of circles that is centred at the receiving position. The network density perceived by a sensor node used in detecting forest fire provided an accurate approximation of forest fire but did not consider the rapid spread of the fire and the memory required by each sensor node. Further research on wildfire detection combined the use of image clustering and wireless sensor network data to train algorithms to detect wildfire. Divya *et al.*, (2014) proposed the use of image clustering to predict the occurrence of wildfire while introducing a mechanism to ensure secure transmission of data from the WSN to prevent threats due to malicious nodes. They were able to categorize wildfire based on the risk and confidence level. However, this approach depends on collection wildfire images which could be a daunting task and the burned area of the existing fire is not considered only the severity of the fire in terms of temperature. Giglio, Schroeder, and Justice, (2016) research provided an insight on the limitations previously observed in the use of MODIS collection 6 as analysis on the update provided by the results of this research indicated that improvements in the emission of errors over large fires reducing the alarm rates in a tropical ecosystem.

2.2 Machine Learning Approach

The machine learning approach extends the traditional approach by training these algorithms from the data focused on improving the detection accuracy for wildfire predictions.
The machine learning approach was applied using variables that show fire activities, geographic locations and metrological attributes to predict the minimum height of smoke and explaining 80% of the variability in the model Yao *et al.*, (2018). They criticized random forest algorithm to be limited when compared to linear logistic regression but offers good predictive power for analysis and attributed the near-real-time meteorological data they utilized to the reason for the high predictive power. The Random forest algorithm accuracy was further investigated for wildfire severity and found to perform better than the traditional approach of Normalized burned Ratio (NBR) when applied to Landsat satellite platforms but autocorrelation in the features of the data may lead to an inflated accuracy in the algorithm and needs to be accounted for during the analysis of the model. Himawari-8 geostationary satellite data for detecting wildfire in South Korea using random forest and threshold-based algorithm to monitor the high rate of false alarms of wildfire in the region (Jang *et al.*, 2019). The analysis was able to ascertain whether an alarm raised for wildfire occurrence is false or not in 10 minutes, this time is too much for the wildfire to cause a hazard to the environment and resources could be wasted if deployed on false alarm thereby the reason behind the research question. Random forest is part of the machine learning algorithm that would be employed to train the proposed model with the improved accuracy presented in the research reviewed above while checking for autocorrelation to make sure the accuracy of the model will not be inflated by the autocorrelation of the different attributes in our model and are difficult to implement with large dataset

Most wildfire prediction systems contain many monitor features and are very difficult to implement in developing countries witnessing such natural phenomena like wildfire. Mhawej, Faour and Adjizian-Gerard (2015) identified climate, topographi, in-situ, historical

and anthropogenic as factors related to ignition of wildfire and can be used for better assessment of wildfire risk. To control for a large number of features in the analysis based on artificial neural network (ANN) and Support Vector Machine (SVM) algorithms showing an improved error rate in analyzing wildfire predictions (Sakr, Elhajj and Mitri, 2011). They are the wildly used machine learning algorithms and SVM performed better than ANN when classifying the occurrence of fire while the error rate of ANN was lower than SVM on average. The algorithm is found to have a high R-squared value and was able to predict wildfire without computation on the whole dataset increasing the near real-time process of wildfire detection relating back to the research question proposed by this study. The lower the RMSE and mean standard deviation (MAD) the better the result but the RMSE is exposed to high errors which can also be compared to the test error. The universal formula used for calculating the performance of the model in terms of RMSE and MAD is stated in the equation 1 and 2 (Kansal *et al.*, 2016);

$$MAD = \frac{1}{k} * \sum_{i=1}^{k} |b_i - \hat{b_i}| \qquad (1)$$

$$RMSE = \sqrt{\sum_{i=1}^{k} (b_i - \hat{b_i})^2 / k} \qquad (2)$$

The metric is given as Mean Absolute Deviation (MAD) and Root Mean Squared Error (RMSE), the lower values are discovered to yield better results but the RMSE is sensitive to high error. Rodrigues and De la Riva, (2014) surveyed human-induced wildfire with machine learning models applying Random Forest (RF), Boosting Regression Trees (BRT), Support Vector Machine (SVM) accessing the models using linear regression (LR). SVM yielded higher accuracy than the decision tree counterparts, the model was found to be inadequate when classifying wildfire as computation of the calibrations was found to be time-consuming which is the part of the research question presented in this research to effectively utilize distributed systems in instances like this. In contrast to the model built by Rodrigues & De la Riva, (2014), Kansal *et al*., (2016) argued that regression used to evaluate the models performed better in wildfire detection by splitting the dataset which they stated will work best and had better accuracy compared to the other machine learning models. This is limiting for a larger dataset and contains bias tendencies as the whole aspect of the research is not considered when the data is split as referring the computation of the large data that need to be considered to effectively carry out predictions. Ralph and Carvel, (2018) proposed a hybrid model for building these fire safety models pointing out the lack of experimental validation is critically lacking for models that make use of these hybrid models. The issue of model weakness and choice of model to use when building the safety model but this hybrid coupled model will provide an effective way of building these safety models. In machine learning, an algorithm is considered superior over another one when a through experimental results accompanied by statistical analysis are conducted and the difference between models is

significant (Settouti, Bechar and Chikh, 2016). Settouti, Bechar, and Chikh, (2016) ranked C4.5, Bagging, Adaboost, CART and SVM as the top five most algorithms to consider when performing machine learning analysis.

2.3 Study Area and Input Data

This subsection reviews the study area and input data considered by different researchers relevant to this study area and domain. The study area and type of input data considered is an important determinant factor as studies have shown that land-cover class fire behaves selectively. Oliveira *et al*, (2014) found that in Europe wildfire is less likely to occur in artificial surfaces and agricultural areas but wildfire was significantly large in shrublands and grasslands. This can be attributed to the investigation of Ganteaume *et al.*, (2013) on ground fuel flammability and its ability to cause a fire, pointing out that the higher density and moisture content of carburant can increase the time until ignition of fire can be observed. They concluded that an increase in the wildfire is as a result of negligence with the current land-use cover change and Abiotic factors related to weather, fuel, and topography are significant environmental factors driving the ignition of wildfire.

Radke, Hessler, and Ellsworth, (2019) based their research on GeoMAC data from geographic information systems (GIS) considering the United States as a case study. The research was conducted using limited parameters but leveraged the use of deep learning to create a firecast model able to identify high-risk areas of fire spread up to 2 weeks as the current state of the art wildfire spread models relied on mathematical growth predictions and physics growth model. Srivas *et al.*, (2016) extended the software used for modeling and simulation of wildfire to include data assimilation from noisy and limited spatial resolution to improve the accuracy of the wildfire predictive model. This research utilized the weather data from California in the United States results indicating the ability of the model to converge on the actual wildfire perimeter in only a few data assimilation steps but include large when these steps are not properly performed making the model prone to errors. Mithal *et al.*, (2018) applied machine learning framework on the tropical forest of South America and South-east Asia with a combination of two approaches able to provide a comprehensive assessment of the tropical fire. However, the chance of having considerable variability with the MODIS tile and the MODIS land cover labels may be incorrect thereby impacting the accuracy to identify fires.

Stojanova *et al*., (2006) developed predictive models from Geographic information System (GIS) data, Meteorological ALADIN (Aire Limitée Adaptation dynamique Développement InterNational) data and MODIS data. In other to cover the different climatic conditions of Primorska region, Kras region of western Slovenia and the continental part of Slovenia, they considered several features that are not relevant to the analysis indicating the need for feature selection to extract relevant feature.

2.4 Summary of Literature Survey

The review of recent literature presents the relevance of this research with the building blocks to the novel approach utilized in this research. It is crucial to point out that the recent and

credible research of Jang *et al.*, (2019) will form the criterion for which research was built. Essentially, the key points to note from this survey is the need for the improved computational time of the predictive model, generalized study area in terms of the region and also being able to consider a substantial amount of data to cover every possible aspect of the analysis perform while maintaining a minimum training time. Thus, this research with its novelty approach seeks to attain a significant result in the enhancement of wildfire predictive models as well as the computational time required to train these models

3 Research Methodology

This section is dedicated to present the methodology of the research, established to ascertain if developing unsupervised and supervised learning approaches to minimize the cost and time to respond to wildfire by setting up equipment close to where the fire is likely to occur and computing the probability of its occurrence. This section describes the entire experiment including the procedures, data collection, setup, and the various algorithms used that contributed to different choices and motivations made to answer the research question.

3.1 Data Requirement and Gathering

The dataset used for this project was sourced from National Aeronautics and Space Administration (NASA)[1] for Canadian Region from 2017- 2018. The data sourced was originally used in Cortez and Morais, (2007) research however, the updated data contained 14 attributes with 98712 observations using the Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies- Collection 6 representing the center of a 1km pixel that is flagged by the MODIS. Some of the attributes of the dataset include latitude, longitude, brightness, scan, track, satellite, confidence, day-night. These attributes are related to metrological factors influencing wildfire extracted from Terra and Aqua satellites. After the data has been collected, several preprocessing tasks were carried out, all the missing values removed to improve the performance of the models, categorical variables were properly encoded to correct imperfections on the data.

3.2 Process Flow

The process flow is a step by step procedure followed to implement the solution for this research. This subsection describes specifically the holistic view of the research procedures and customized knowledge discovery and data mining (KDD) approaches are developed in the steps as shown in figure 1.

[1] https://firms.modaps.eosdis.nasa.gov/

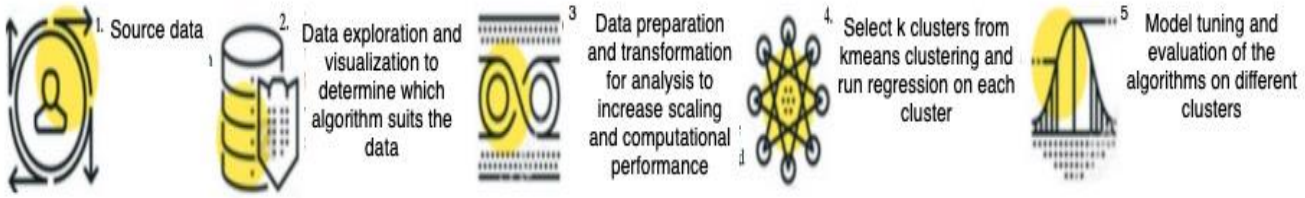


Figure 1: Process flow diagram for the research

In figure 1, the process flow represents the processes involved from sourcing the data to the evaluation of results using k-means clustering and regression algorithms. All these stages (steps 1 to 5) are representative of all the steps carried out in this research and further given context in section 5.

3.3 Overview of Techniques and Data Mining Algorithms Implemented

This subsection presents a detailed description of the data mining algorithms and statistical methods deployed for this research and the reasons for choosing these methods. For this research, three statistical methods and three data mining algorithms were achieved in this research. These data mining algorithms and statistical methods include:

- Kmeans clustering - is an unsupervised learning approach used in this research. The algorithm is an iterative algorithm that tries to partition the dataset into k-number of centroids that are distinct non-overlapping clusters allocating each variable to their nearest cluster to only one group. K-means clustering is considered the best and simplest clustering technique for this type of dataset.
- Cook's distance - this statistical technique is used to identify influential data points in the dataset. This technique calculates the residual values of all the variables to identify influential data points and resolving these data points to remove bias from the analysis
- Pearson correlation - Pearson correlation gives the correlation between -1 and 1, a Pearson correlation value of -1 and 1 signifies high correlation. For this research, multicollinearity was avoided in the regression algorithm to achieve a non-biased result for the model.
- Silhouette Analysis - this validation technique is used in this research to determine the degree of separation between the clusters and a value close to 1 signifies a good cluster.
- Random forest – this machine-learning algorithm uses the bagging technique to sample the training data by constructing decision trees and outputting the class that is the mode of the classes of the individual tree. (Collins *et al.*, 2018)utilized random forest (RF) to improve the accuracy of satellite-based wildfire severity mapping using Landsat imagery but this research will be deployed for regression and identifying the variable that is important to predict the occurrence of wildfire.
- Cross-validation – this validation technique is used to investigate the authenticity of the result, meaning the result was not due to chance. The cross-validation is

implemented to validate the classification and regression technique to ensure a reduced error rate using 5 iterations

- Synthetic minority over-sampling (SMOTE) – this technique is used to balance the minority class during classification. Yao *et al.*, (2018) established in their research the importance of a fair balance dataset and its importance to results obtained. For this research, the SMOTE is used to oversample the minority class and using more training data during the model training to prevent loss of information and achieve improved results.
- Navie bayes – this is a machine learning technique utilized in simple multiclass classification with the assumption of independence between every pair of features. Cortez and Morais, (2007) applied this algorithm to determine the burned area of a wildfire in their research and achieve a significant result. This algorithm computes the conditional probability distribution of each label feature give and use it for prediction

4 Design Specification

This section describes the implementation architecture adopted for this research. The 3-Tier architecture is implemented in this research as proposed by Sayad, Mousannif and Al Moatassime, (2019) which is flexible, reusable and is well suited for the analysis carried out in this research. The 3-Tier architecture deployed in this research is shown and described in Figure 2.

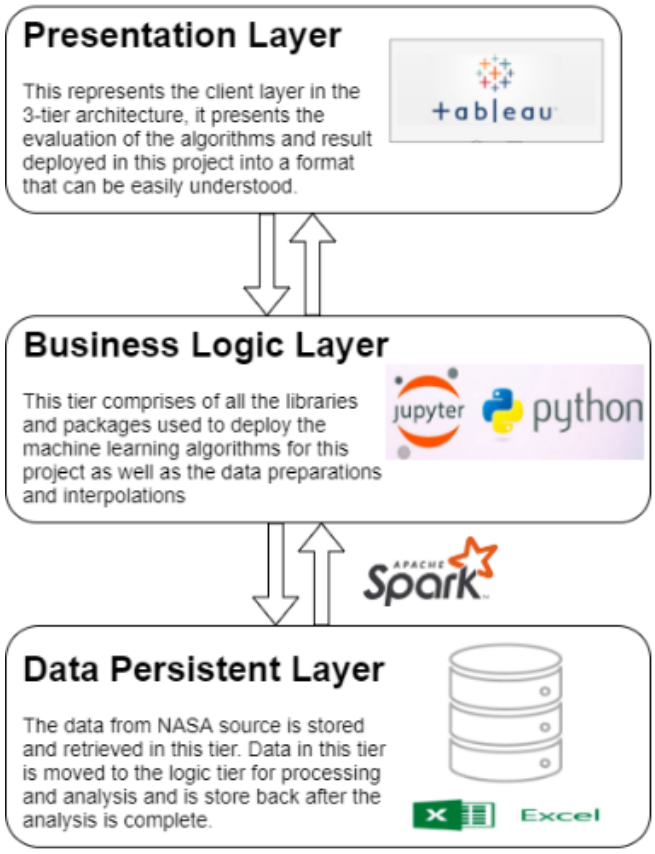


Figure 2: A 3-Tier Architectural Design.

5 Implementation

This section describes and visualizes the final stages and steps of the research methodology.

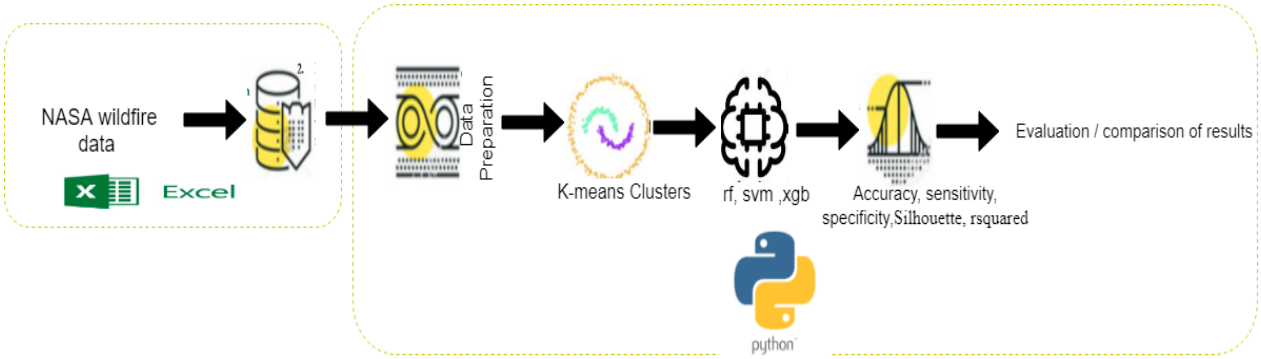


Figure 3: Implementation of wildfire predictive model.

The implementation of this research as shown in figure 3 done on the NASA wildfire dataset based on Cortez and Morais (2007) research extracted from NASA fire archive and placed in a csv format and pre-processed using python. The k-means cluster is used to select the best natural fire cluster to pick out central locations of each of these clusters to enable fire equipment and resources placed at the closest fire active points. For regression, the k-means clusters are evaluated, selecting useful features and applying a machine learning pipeline using Random forest and navie bayes on the data. The Silhouette evaluation analysis is used to establish the distance of each of the clusters from the centroid point while RMSE and R-squared are established and evaluated to identify how well the model performed. The next subsection will discuss the diagram represented in figure 3 in context.

5.1 Data Pre-processing and Exploration

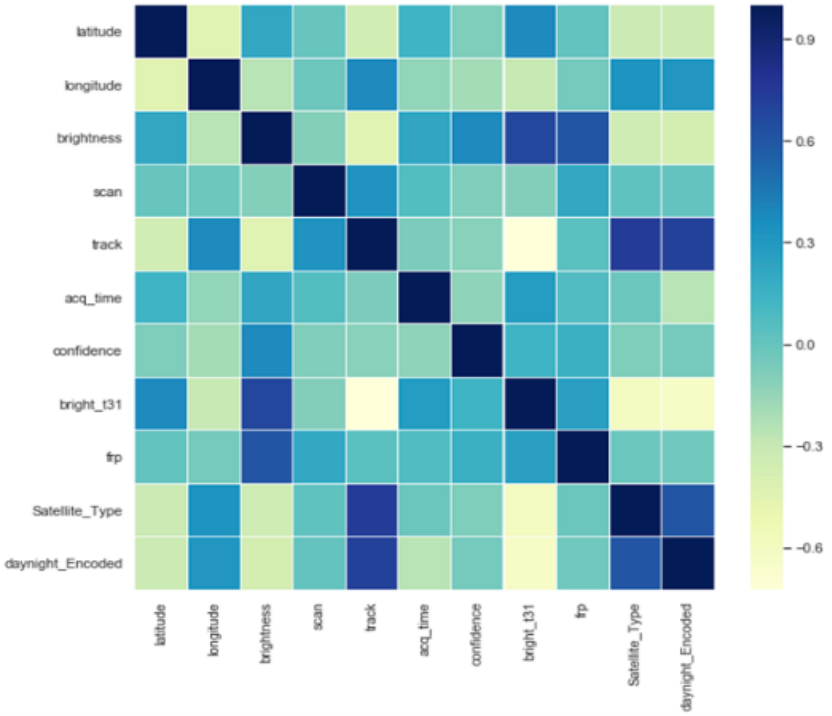
The first step in the data preprocessing is understanding the features of the data which consists of applying various feature engineering techniques to regularize the dataset. This subsection describes all the quality checks and exploration in the dataset.

The wildfire data in Canadian regions from 2017 to 2018 sourced from the NASA archive database containing 98712 observations with 15 wildfire indicators. In dealing with data, checks for missing values shows that there were no missing values from the structured raw data collected from the NASA archive download as shown in figure 4A, the presence of multicollinearity in each cluster where represent in a heatmap as shown in figure 4B and a descriptive of the correlation shown in figure 4C. The non-numeric variable in the dataset was cast to float enabling the computational analysis to be done on the data.

A

	latitude	longitude	brightness	scan	track	acq_time	confidence	bright_t31
count	98888.000000	98888.000000	98888.000000	98888.000000	98888.000000	98888.000000	98888.000000	98888.000000
mean	55.660454	-116.434807	333.005096	1.656464	1.257596	1529.967041	66.307930	293.139587
std	5.979945	13.066771	32.534572	0.856002	0.827215	655.083069	30.967909	16.438375
min	1.000000	-141.368393	0.740000	1.000000	1.000000	0.960000	0.000000	0.480000
25%	51.707876	-123.709424	313.899994	1.100000	1.000000	812.000000	48.000000	287.200012
50%	53.198098	-120.315647	325.799988	1.300000	1.100000	1912.000000	74.000000	294.100006
75%	60.559399	-109.189877	343.899994	1.900000	1.400000	2030.000000	94.000000	299.899994
max	68.532997	14.830000	505.799988	4.800000	30.000000	2343.000000	100.000000	400.100006

B



C

	latitude	longitude	brightness	scan	track	acq_time	confidence	bright_t31	frp	Satellite_Type
latitude	1.000000	-0.435687	0.222576	0.007950	-0.349092	0.140970	-0.071355	0.384753	0.019193	-0.308674
longitude	-0.435687	1.000000	-0.250479	-0.021024	0.380874	-0.132527	-0.190050	-0.288473	-0.042756	0.338150
brightness	0.222576	-0.250479	1.000000	-0.093588	-0.427767	0.226909	0.382122	0.684307	0.613993	-0.331177
scan	0.007950	-0.021024	-0.093588	1.000000	0.340567	0.073650	-0.080828	-0.088225	0.215534	0.027688
track	-0.349092	0.380874	-0.427767	0.340567	1.000000	-0.068009	-0.107800	-0.724005	0.049905	0.739743
acq_time	0.140970	-0.132527	0.226909	0.073650	-0.068009	1.000000	-0.118187	0.288568	0.082377	-0.004282
confidence	-0.071355	-0.190050	0.382122	-0.080828	-0.107800	-0.118187	1.000000	0.144304	0.168207	-0.077432
bright_t31	0.384753	-0.288473	0.684307	-0.088225	-0.724005	0.288568	0.144304	1.000000	0.269309	-0.585661
frp	0.019193	-0.042756	0.613993	0.215534	0.049905	0.082377	0.168207	0.269309	1.000000	-0.008117
Satellite_Type	-0.308674	0.338150	-0.331177	0.027688	0.739743	-0.004282	-0.077432	-0.585661	-0.008117	1.000000

Figure 4: A) Descriptive statistics; B) Heat map of correlation; c) Correlation descriptive

The boxplot on the brightness indicator showed the presence of an outlier which can lead to bias in the result and inaccurate conclusion. Additional investigation using the cooks' distance showed that the value is significantly distanced apart from others increasing the effect on decision making. This value was excluded for the dataset also the data collected using the MODIS instrument as already been geo-referenced and not distorted, but the values in the dataset may be incorrect due to wrong energy reading from the sensor requiring radiometric correction (Sayad, Mousannif and Al Moatassime, 2019).

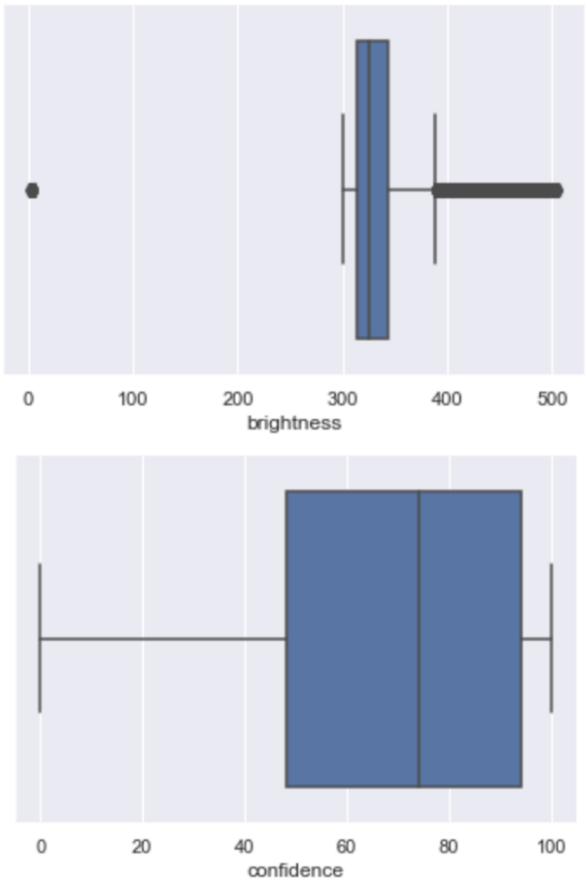


Figure 5: Presence of outliers

Further data exploration was conducted on the data using the tableau to further gain an understanding of the data. The data visualizations showed that the bulk of the vegetation fire recorded in the dataset occurred during the summer and autumn period also with a high temperature reading from different temperature instruments used for measurement. Using the longitude and latitude indicator in the dataset was used to investigate fire pixels and radiative power in Canadian regions. The result of the visualization shown in figure 6, shows that cities like Winnipeg, Ontario, Manitoba, Alberta, and British Columbia experienced a high volume of wildfire during these seasons. Cortez and Morais. (2007) carried out their research to predict the burned area of wildfire using Alberta as a case study however there is also a need

to consider other regions thereby supporting the motive of this research to mitigate the risk of wildfire risk.

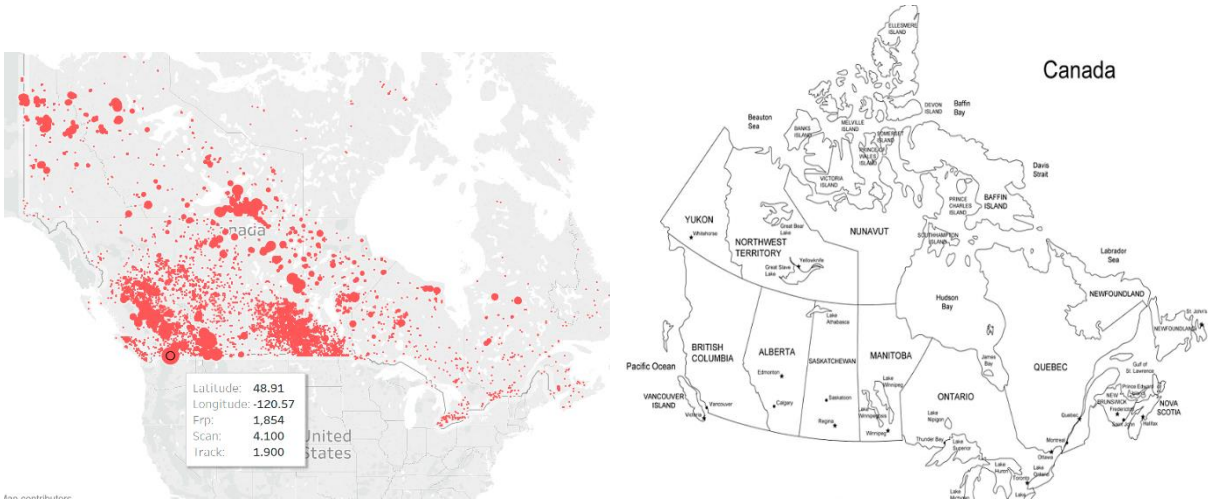


Figure 6: Fire pixel distribution in Canadian Regions

The variables were reduced to 13 variables after performing feature engineering using the Boruta library with a maximum run of 500 to determine important indicators. The confidence indicator ranging between 0 and 100 percent is used to gauge the quality of fire pixel was binned into 3 categories, from 0 to 30% representing “low-confidence fire”, 31% to 80 % representing “nominal-confidence fire” while from 80% and above represented “high-confidence fire”. These categories were determined by thresholding the confidence value calculated for the fire pixel as shown in table 1. The low-confidence fire pixel was considered for this analysis to get maximum detectability of fire but at the cost of tolerating higher incidence of false alarms in the analysis.

Range	Confidence Class
$0\% \leq C < 30\%$	low
$30\% \leq C < 80\%$	nominal
$80\% \leq C \leq 100\%$	high

Table 1 Fire pixel confidence class (Giglio *et al*, 2018)

The data binning of fire pixel into categories caused the dataset to be imbalanced as shown in figure 6 (A), applying SMOTE to the dataset as shown in figure 6 (B) tends to balance the minority classes in the dataset. Using the SMOTE technique presents the risk of overfitting during training and to prevent this the number of training samples was increased to capture this effect on the dataset.

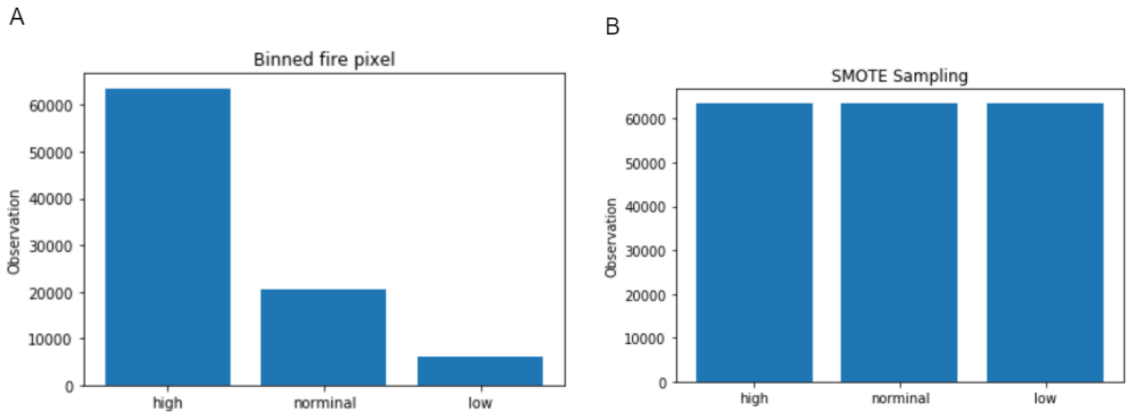


Figure 7: A) Class distribution of imbalanced data; B) Oversampling the minority class

5.2 Application of Data Mining Algorithms

This section describes the implementation of four different data mining algorithms applied at different stages of the experiment.

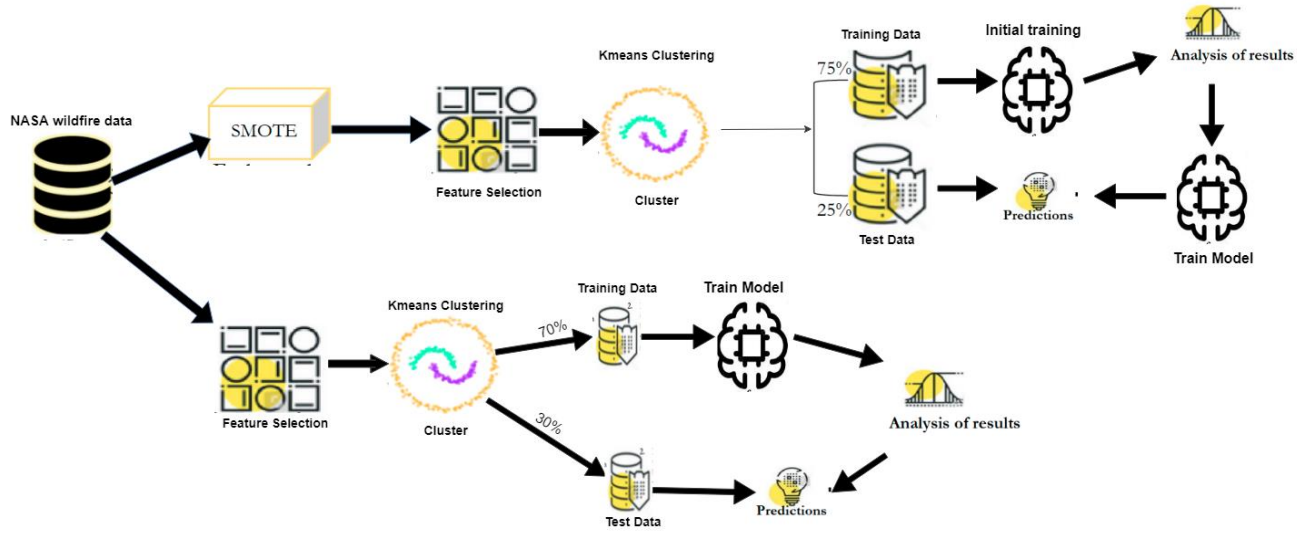


Figure 8: Overview of wildfire predictive model

The data was split into 70% as a training sample from the dataset and the remaining 30% for the test data during initial analysis with the imbalance dataset before applying SMOTE to balance the minority class for classification. After applying the SMOTE to balance the minority class, the data was split to 80% for training and 20% for testing to gain a better understanding of the predictive model and using cross-validation to validate the results obtained during classification in terms of accuracy specificity and sensitivity. The cross-validation technique is used to check for the best tuning model using RMSE and this tuning is applied to predict wildfire and establish the importance of various indicators used in building the model. The algorithms implemented in the research is as established above in subsection

3.3 using Pearson correlation to ensure the absence of multicollinearity in the independent variables.

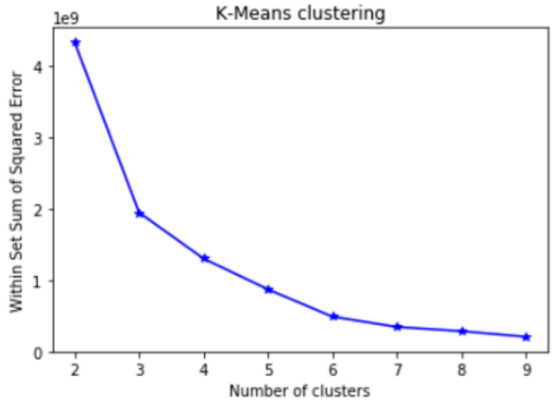
6 Evaluation

In data mining, performance evaluation is based on the type of problem resolved and the results obtained. Therefore, the validation technique implemented for this research was chosen to enable the research question and the literature review. The results and evaluations carried out in this research are based on the literature review as well as accepted performance metrics for regression and classification deployed in this project. The RMSE value of zero represents the absence of error from the prediction while for the R-squared, a value of one show how well the model explained the variation of the dependent variable. This section provides a comprehensive analysis of the results of different stages in the experiment building up to this research.

6.1 Experiment 1 (Kmeans Clustering)

The first experiment for this research is the K-means clustering of the dataset, to determine the number of clusters for the dataset the elbow method was applied to the dataset resulting in the diagram shown in figure 9A. To find the best value of k using the silhouette plot as shown in figure 9B was used to measure how similar each cluster is compared to another cluster with the algorithm correctly clustering 70% of the dataset to its appropriate cluster. Figure 9C shows the resulting visualization of the 4 clusters used for this analysis and their respective centroids.

A



B

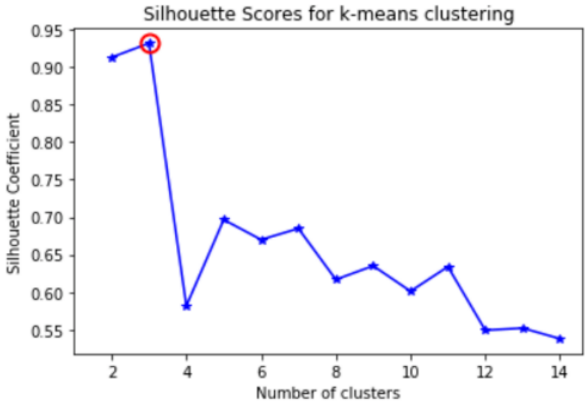


Figure 9: A) K-means elbow method; B B) Silhouette score analysis;

C

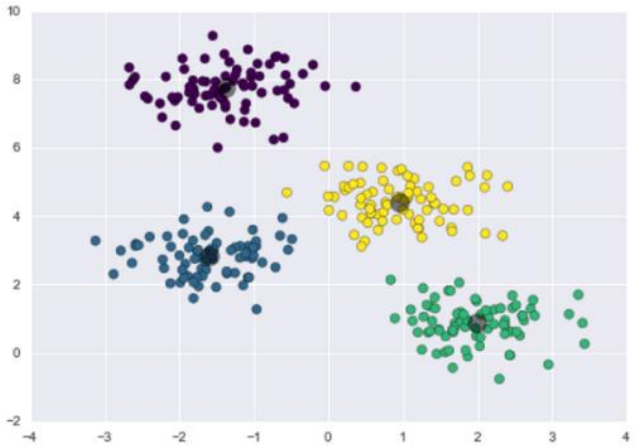


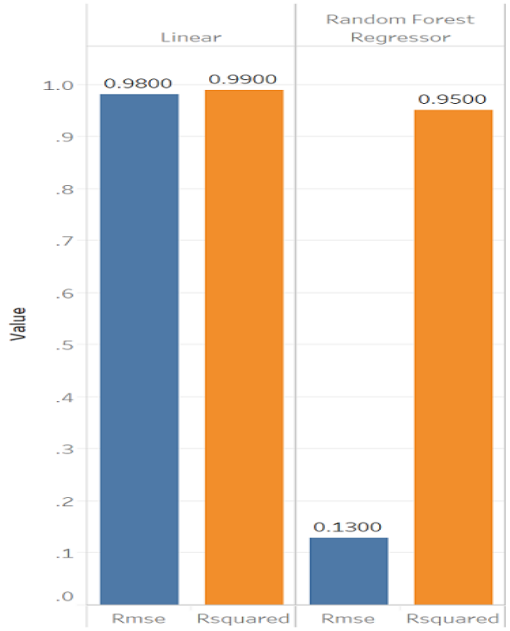
Figure 9: C) K-means Cluster plot

6.2 Experiment 2 (Regression (Linear and Random Forest)

The second experiment conducted was applying regression on the clustered dataset with elastic net parameter and regular parameter. Cross-validation of the model was performed to determine the best performing variable reporting the values for each variable. Figure 10A

shows the distribution of RMSE and R-squared for the linear and random forest algorithms applied to the clustered data, indicating that linear regression performed better compared to the random forest regressor in terms of RMSE and Rsquared. Figure 10B describes the cross-validation results of the clustered dataset, with results indicating that confidence value of fire pixel denoted as "confidence" was the best performing variable controlling for 78% of the model.

A



B

	values	indexedFeatures
0	0.001490	latitude
1	0.002477	longitude
2	0.144992	brightness
3	0.002491	scan
4	0.002210	track
5	0.003916	acq_time
6	0.779772	confidence
7	0.012713	bright_t31
8	0.045486	frp
9	0.001704	Satellite_Type
10	0.002750	daynight_Encoded

Figure 10: A) Distribution of RMSE and R-squared values; B) Cross validation result for the model

6.3 Experiment 3 (Classification (Naïve Bayes and Random Forest)

The third experiment carried out was the application of classification algorithms to predict the binned confidence for fire pixels in the dataset. The binned confidence as the dependent variable had an imbalanced class therefore, the algorithm was analyzed in two stages. The first stage was applying the algorithm on the imbalance data while the second stage was applying the algorithm on balanced using SMOTE to oversample the minority class. The accuracy, specificity and sensitivity and precision are reported in each stage of the analysis. The confusion matrix of the navie bayes classifier for the imbalance phase is shown in figure 11 with the model correctly classifying 10037 out of 12593 "high fire pixel", 3988 out of 4050 "nominal fire pixel" and 1158 out of 1254 "low fire pixel". The normalize and non-normalized confusion matrix for the navie bayes classifier is shown in figure 11.

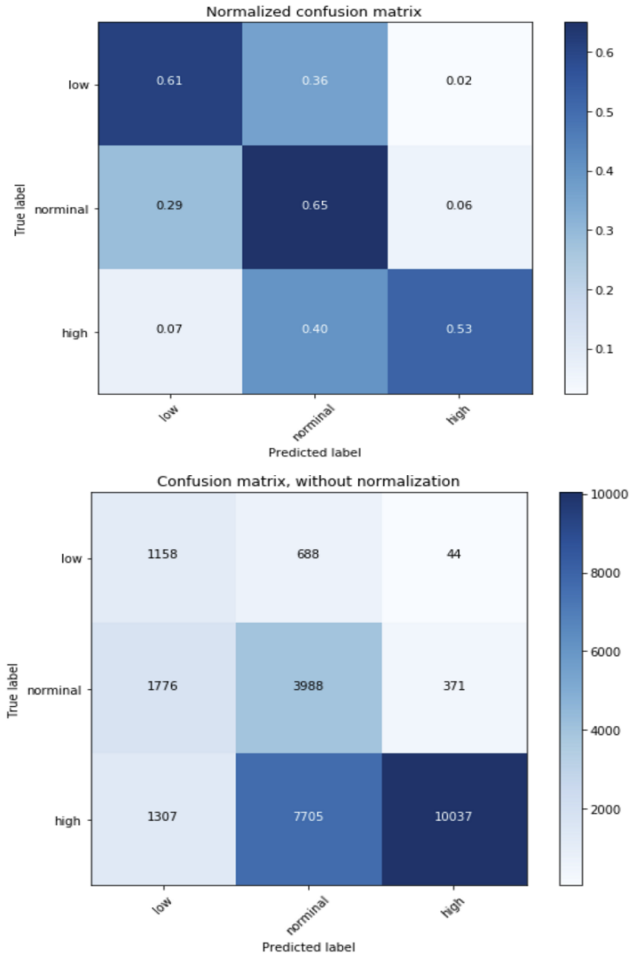


Figure 11: Confusion matrix for imbalance naïve bayes model

Contrary results were observed applying SMOTE on the dataset for the oversampling phase as no significant change was observe as shown in figure 12 for the evaluation metrices used to validate the model. The result from the SMOTE sampling also shown in figure 12, that the model correctly classified 9639 for the “high fire pixel”, 12455 for the “nominal fire pixel” and 11713 for the “low fire pixel”. This phase was carried out to ensure that the result obtained was not by chance and the normalize and non-normalized confusion matrix for navie bayes classifier using SMOTE is shown in figure 12

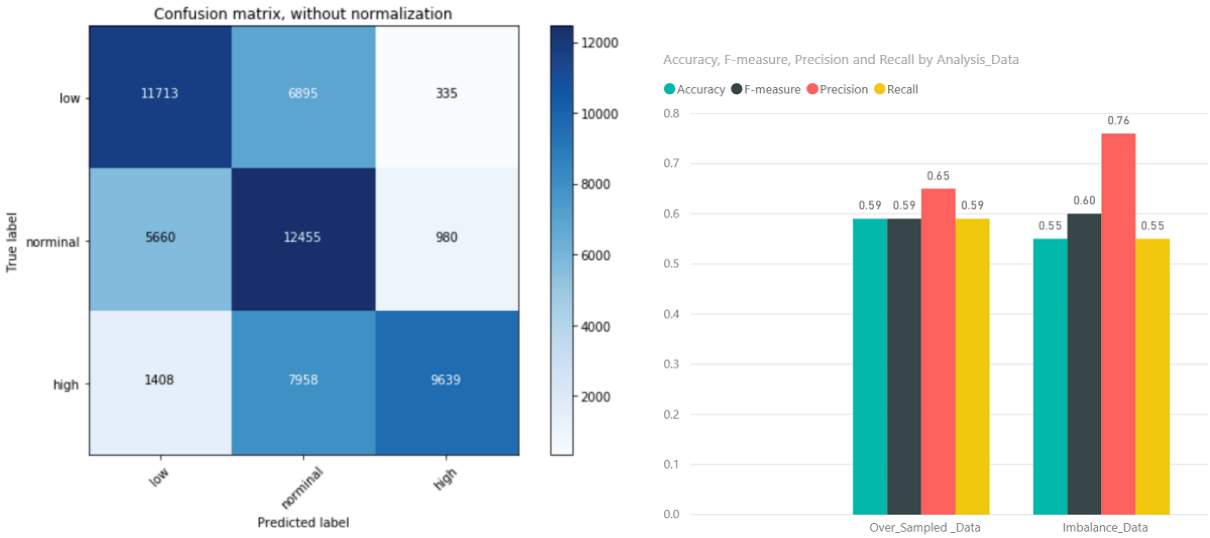


Figure 12: Confusion matrix for SMOTE naïve bayes model

For the Random forest as shown in figure 13, the confusion matrix on the right shows the imbalance phase was observed to classify records belonging to each of the subclasses without any predicted value among them while that on the left shows the result on applying oversampling to the dataset. Figure 13 shows the normalized and non-normalized confusion matrix for the classifier. The performance metrics for the accuracy, sensitivity, and sphericity for the imbalanced phase was 96% for each of the performance metrics while for the SMOTE, the value for each of the performance metrics increased to 97% and the model was able to predict classify other classes in the dataset.

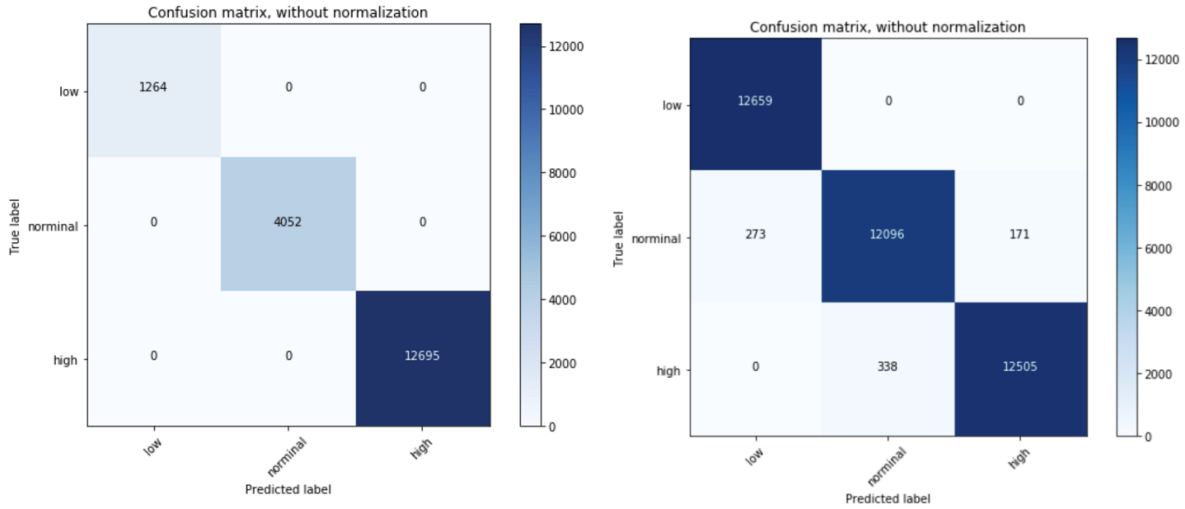


Figure 13: Confusion Matrix of Random forest (imbalance SMOTE)

6.4 Discussion

The results obtained from the 3 experiments conducted for this research is explained in this section as well as the interpretation of result from other research expressed using classification metrics.

The algorithms used in this research are supervised learning and unsupervised learning (Settouti, Bechar and Chikh, 2016) based on the research question this research is seeking to address. The unsupervised machine learning was analyzed because it could aid the fire mangers to build an effective fire response model from the K-means cluster which has proven to use in discovering patterns in data while the supervised learning was applied for the classification and regression in the research.

In the first experiment, the k-means for two clusters had the best silhouette percentage of 94% as shown in the figure 9B however for this research made use of four clusters based on the evaluation of the elbow method depicted in figure 9A as clusters above four tend to have a constant silhouette percentage. The second experiment was the application of regression and as earlier established, RMSE of 0 indicates the absence of error in the analysis while the R-squared value of 1 depicts a perfectly explained model. The error rate in the linear regression was 0.99 while the random forest regressor had a reduced RMSE of 0.13 while the R-squared value for both the algorithms significantly explained the model. An investigation showed that the random forest regressor proved its usefulness and strength for regression with a desired reduced error rate.

For the third experiment, random forest and navie bayes were considered based on the model previously used by researchers in this domain (Cortez and Morais, 2007). These algorithms were applied using distributed computing platform Apache spark, but the choice of algorithm for classification was limited as a result of the type of analysis to be conducted and conflicts on updates from the spark 2 machine learning library as at the time of this research with improper documentation. However, the computational time using the distributed computing platform was impressive as the navie bayes took 4 seconds to train 72239 samples while the random forest training time was 7 seconds on the same number of samples. The analysis for classification models was in two phases, the first phase comprised of implementing both algorithms using an imbalanced class set with the “high fire pixel” class in the majority while the second phase utilized oversampling of the class set using SMOTE. The distribution of the performance metric for both models as shown in figure 14 and no significant result was observed applying oversampling on the dataset. Performance metrics alone do not define how good a model is, therefore, other validation metrics are required (Sayad, Mousannif and Al Moatassime, 2019). Cross-validation was applied with hyperparameter turning on the training models and no significant value increase or decrease was observed in the model supporting obtained in the research.

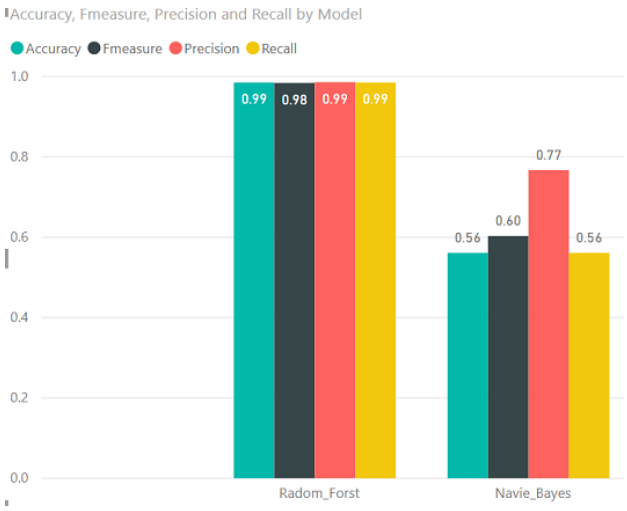


Figure 14: Performance metrics of algorithms for classification.

Overall, random forest outperforms other models used in this research for regression and classification analysis in terms of performance but in computational time, navie bayes was faster in model prediction. The experiment was set in a bid to fulfil the set objectives for this research and enable the research question to be answered. Thus, this research concludes with a combination of K-means clustering, and random forest deployed on distributed computing framework as having the predictive performance, decrease computational time with reduced error rate in the prediction of wildfire.

7 Conclusion and Future Work

In this paper, the aim is to analyze wildfire predictive models using distributed computing as it threatens the lives of people in Canadian regions. Over the years, thousands of hectares of land in the forest are lost to wildfire with billions of Canadian dollars spent on resources to reduce the effect on structure, composition and soil fertility of the forest. Hence to mitigate wildfire risk, this research analyzed the wildfire predictive models implemented on a distributed computing framework "Apache spark" making use of dataset from satellite images using Moderate Resolution Imaging Spectroradiometer (MODIS).

The extracted data from the National Aeronautics and Space Administration (NASA) was preprocessed and analyzed using supervised and unsupervised machine learning algorithms. Random forest, navie bayes, and linear regression are supervised machine learning models used for this research while K-means clustering was used for the unsupervised machine learning model. The model gave good results for regression with the RMSE values less than that of linear and random forest regressors for analysis while the value of R-squared in random forest regressor was lesser compared to the linear regressor. For classification, the accuracy was assessed using classification metrics and cross-validation for both algorithms with random forest having an average of 97% across sensitivity, specificity, and accuracy confirming the efficiency of the model in predicting wildfires in the least amount of time thereby answering the research question posed in this research.

This research effectively analyzed wildfire predictive models introducing distributed computing that can greatly improve the computational time in building predictive models in this domain. However, there may be considerable amount of false alarm within the low-confidence fire pixel class that can influence the ability of the model to identify fires with high accuracy. The scope of this paper could be extended by not considering the low-confidence fire-pixel in the analysis and utilizing more remote sensing data indicators which can have an impact in predicting wildfire such as soil vegetation, humidity and wind speed of the study area to be considered. This research could also be extended by applying other algorithms not considered in the scope of this study to this domain using a distributed computing platform.

Acknowledgement

I acknowledge the use of data and imagery from LANCE FIRMS operated by NASA's Earth Science Data and Information System (ESDIS) with funding provided by NASA Headquarters.I would like to thank my family and friends for the support I received throughout my academic study. Finally, I would like to thank the Irish Government for allowing me to study data analytics at the National College of Ireland.

References

Collins, L., Griffioen, P., Newell, G., & Mellor, A. (2018). The utility of Random Forests for wildfire severity mapping. *Remote Sensing of Environment, 216*(July), 374-384.

Cortez, P., & Morais, A. (2007). A Data Mining Approach to Predict Forest Fires using Meteorological Data. *Proceedings of the 13th Portuguese Conference on Artificial Intelligence, 2014*(January 2007), 512-523.

Divya, T., Manjuprasad, B., Vijayajakshmlm, N., & Dharani, A. (2014). An Efficient and Optimal Clustering Algorithm For Real-Time Forest Fire Prediction with Sensor Networks and Data Mining. *2014 International Conference on Communication and Signal Processing*, 312-316.

Ganteaume, A., Camia, A., Jappiot, M., San-Miguel-Ayanz, J., Long-Fournel, M., & Lampin, C. (2013). A review of the main driving factors of forest fire ignition over Europe. *Environmental Management, 51*(3), 651-662.

Giglio, L., Schroeder, W., & Justice, C. (2016). The collection 6 MODIS active fire detection algorithm and fire products. *Remote Sensing of Environment, 178*, 31-41.

Jang, E., Kang, Y., Im, J., Lee, D.-W., Yoon, J., & Kim, S.-K. (2019). Detection and Monitoring of Forest Fires Using Himawari-8 Geostationary Satellite Data in South Korea. *Remote Sensing, 11*(3), 271.

Kansal, A., Singh, Y., Kumar, N., & Mohindru, V. (2016). Detection of forest fires using machine learning technique: A perspective. *Proceedings of 2015 3rd International Conference on Image Information Processing, ICIIP 2015*, 241-245.

Mhawej, M., Faour, G., & Adjizian-Gerard, J. (2015). Wildfire Likelihood's Elements: A Literature Review. *Challenges, 6*(2), 282-293.

Mithal, V., Nayak, G., Khandelwal, A., Kumar, V., Nemani, R., & Oza, N. (2018). Mapping burned areas in tropical forests using a novel machine learning framework. *Remote Sensing, 10*(1), 1-23.

Radke, D., Hessler, A., & Ellsworth, D. (2019). Firecast: Leveraging deep learning to predict wildfire spread. *IJCAI International Joint Conference on Artificial Intelligence, 2019-Augus*t, 4575-4581.

Ralph, B., & Carvel, R. (2018). Coupled hybrid modelling in fire safety engineering; a literature review. *Fire Safety Journal, 100*(February), 157-170.

Rodrigues, M., & De la Riva, J. (2014). An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling and Software, 57*, 192-201.

Sakr, G., Elhajj, I., & Mitri, G. (2011). Efficient forest fire occurrence prediction for developing countries using two weather parameters. *Engineering Applications of Artificial Intelligence, 24*(5), 888-894.

Sayad, Y., Mousannif, H., & Al Moatassime, H. (2019). Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire Safety Journal, 104*(January), 130-146.

Serna, M., Bermudez, A., & Casado, R. (2013). Circle-based approximation to forest fires with distributed wireless sensor networks. *IEEE Wireless Communications and Networking Conference, WCNC*, 4329-4334.

Settouti, N., Bechar, M., & Chikh, M. (2016). Statistical Comparisons of the Top 10 Algorithms in Data Mining for Classi cation Task. *International Journal of Interactive Multimedia and Artificial Intelligence, 4*(1), 46.

Singh, Y., Saha, S., Chugh, U., & Gupta, C. (2013). Distributed event detection in wireless sensor networks for forest fires. *Proceedings - UKSim 15th International Conference on Computer Modelling and Simulation, UKSim 2013*, 634-639.

Srivas, T., Artés, T., De Callafon, R., & Altintas, I. (2016). Wildfire spread prediction and assimilation for FARSITE using ensemble kalman filtering. *Procedia Computer Science, 80*, 897-908.

Stojanova, D., Kobler, A., Džeroski, S., & Taškova, K. (2006). Learning To Predict Forest Fires. *In 9th International multiconference Information Society*(January), 3-6.

Van Der Aalst, W., Rubin, V., Verbeek, H., Van Dongen, B., Kindler, E., & Günther, C. (2010). Process mining: A two-step approach to balance between underfitting and overfitting. *Software and Systems Modeling, 9*(1), 87-111.

Yao, J., Raffuse, S., Brauer, M., Williamson, G., Bowman, D., Johnston, F., & Henderson, S. (2018). Predicting the minimum height of forest fire smoke within the atmosphere using machine learning and data from the CALIPSO satellite. *Remote Sensing of Environment, 206*(March 2017), 98-106.

Young, J., Thode, A., Huang, C., Ager, A., & Fulé, P. (2019). Strategic application of wildland fire suppression in the southwestern United States. *Journal of Environmental Management, 245*, 504-518.

Giglio, L., Schroeder, W., Hall, J., & Justice, C. (2018). *MODIS Collection 6 Active Fire Product User's Guide Revision B.*

Oliveira,S.,Pereira,J.M.,San-Miguel-Ayanz,J.,Lourenc o,L.,(2014).Exploring the spatial patterns of fire density in Southern Europe using geographically weighted regression. *Appl.Geogr, 51,*143–1578.

MODIS Collection 6 NRT Hotspot / Active Fire Detections MCD14DL.Available on-line https://earthdata.nasa.gov/firms]. doi:10.5067/FIRMS/MODIS/MCD14DL.NRT.006