Wildfire prediction based on Soil moisture using BPNN

MSc Research Project
MSc Data Analytics

Abinandhan Sundar
Student ID: x18113541

School of Computing
National College of Ireland

Supervisor: Sidra Bhasir

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Abinandhan Sundar
Student ID:	x18113541
Programme:	MSc Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Sidra Bhasir
Submission Due Date:	12/08/2019
Project Title:	Wildfire prediction based on Soil moisture using BPNN
Word Count:	5400
Page Count:	20

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:	7th August 2019

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):	

Wildfire prediction based on soil moisture using BPNN

Abinandhan Sundar
x18113541

Abstract

Wildfire is one of the devastating natural disasters, which is partially caused by human and nature. Wildfire prediction can be helpful to save the forest from the fire occurrences. Same is achieved using advanced machine learning and data analytical techniques. The forest fire caused by the human can be controlled, but the naturally occurring wildfire can be controlled to some extinct. The naturally occurring wildfire might cause due to several reasons namely lighting, drought or dry topological changes. The wildfire caused due to lighting is very rare and the same can't be predicted or controlled. The wildfire caused by topological changes namely dead vegetation and surface temperature can be controlled and predicted. The vegetation index is measured as Normalized Difference Vegetation Index (NDVI) and surface temperature is measured as Land Surface Temperature (LST). Weather factors involved in the wildfire are wind, air temperature and soil moisture. Since the wind and the air temperature make the wildfire to spread and choosing the direction of wildfire spreading once the fire is ignited, so these factors will be of no use to this research. Soil moisture is a key factor in understanding the water content of the soil. The crop has higher chances of catching fire in dry soil than the crop in wet soil. This research includes the soil moisture as weather attribute along with mentioned attributes to classify the wildfire class. The state-of-the-art model uses NDVI, LST to predict the wildfire. To strengthen the state-of-the-art model, soil moisture (weather attribute) has been added to understand the is there a statistically significant improvement in prediction accuracy. The final model developed in this research with soil moisture attribute exhibits 99 % of prediction accuracy.

1 Introduction

Wildfires are the very serious problems that threaten to destroy tons of forest resources and thousands of square kilometres of forest areas each and every year. The wildfire is categorized into three major types namely Crown Fires, Surface Fires and Ground Fires. Crown fires is very intense and destroys the forest, which can be caused by lightning or natural calamities. The surface fire occurs rarely. But when the same happens it damages the whole forest. The ground fires is caused by the dead vegetation, weather conditions and surface temperature of the surface (Sayad et al.; 2019). Ground fire occurrences is merely dependent on above said factors, whereas the quantity of dead vegetation leftovers under the ground acts as fuel to fire. The dead vegetation might contain harmful chemical or organic matters which acts has a fuel to catch fire, and without help of a catalyst fuel cannot catch fire on its own. In this case the surface temperature acts as a catalyst to ignite the fire. For the fire to spread further, it depends on the weather factors such as air temperature, wind and soil moisture. This natural calamity can be predicted by means of monitoring all the above factors that causes wildfire using advanced data analysis methods. The factors or fire triangle namely fuel, topology, weather and ignition sources will be collected as data with different scaling factors recommended with the fire and no fire indicator. In Sayad et al. (2019) paper wildfire occurrences can be predicted based on three attributes namely Normalized Differential Vegetation Index (NDVI), Land Surface Temperature (LST), Burned Area/Thermal Anomalies (TA) as independent attributes with the fire class (fire / No fire) labels. The NDVI is an index of dead vegetation content under the ground, LST refers to surface temperature. Since the research doesnt include any of the weather factors, the developed model is incapable of predicting wildfire occurrences which is influenced by weather factors. In Sayad et al. (2019), the author highlights that the existing models can be strengthened by adding weather factors namely wind, air temperature and soil moisture for improving the classification accuracy. As part of this research, weather factors will be added as an additional attribute to the state-of-the-art model of wildfire prediction. Considering the weather factors is a novel idea, but the challenge is which of the various factors of weather influences the occurrence of a wildfire. In Ghahremanloo et al. (2018), the author highlights soil moisture has primary variable among other weather factors for wildfire prediction. The operational and fire management in Australia uses the soil moisture deficit, as a measurement from old empirical water balances model which estimates day to day soil moisture deficit. Also, in Ghahremanloo et al. (2018) among all the other weather factors, soil moisture is considered as key factor which influences the occurrence of wildfires.The soil moisture is included as a primary weather attribute for wildfire prediction and management (Ghahremanloo et al.; 2018). In this research soil moisture will be included as an additional attribute in the state-of-the-art model for wildfire prediction to check is there a significant improvement. This paper has used artificial neural networks to classify the fire and No fire classes. Since the dataset from the paper Sayad et al. (2019) doesn‘t have any information about zone of wildfire occurred on the year 2013 to 2014 [1], whereas estimating the soil moisture for same is impossible. Hence this research will be carried out using own dataset and the data will be extracted from the year 2010 to 2018 extracted data source [2]. British Columbia website will used for labelling the new dataset based on the fire occurrences in British Columbia from the year of 2010 to 2018.

1.1 Research Overview:

This research paper contains following sections.
Literature survey: Examining the past research done on the wildfire classification and probability, selection of data source, components in fire triangle and its part in wildfire occurrences and importance of soil moisture in weather based wildfires.
Methodology: This sections explains the entire work done as part of data extarction to end result and implemented environment.
Design Specifications: This section explains the architecture developed to solve the research questions and the justifications for using the library.
Implementation: This sections describes the steps followed to extract data to develop

[1] https://github.com/ouladsayadyounes/Wildfires

[2] https://modis.ornl.gov/cgi-bin/MODIS/global/subset.pl

the model.
Evaluation: This section contains detailed explanation of results from each model which supported to solve the research questions with output screenshots.

2 Literature Survey

Wildfire imbalances the natural resource ecosystem around the globe by means of mass destruction of wild forest. the same is partially caused by human and mother nature herself by means of dead vegetation, surface temperature and weather factors (Sayad et al.; 2019). Weather factors such as air temperature, wind and Soil moisture needs to be taken into account. The leftovers of the dead vegetation which can be organic or chemical, influences the occurrences of wildfires. There is no evidence that dead vegetation catches fire on its own (Boisram et al.; 2018). The catalyst behind the dead vegetation catching fire is surface temperature (Sayad et al.; 2019), hence this needs to be considered while monitoring and predicting the wildfire occurrences.

2.1 Wildfire occurrence probability and class classification:

Massive destruction to the forest ecosystem is caused by ground fires. With the recent advancements in machine learning and Artificial intelligence wildfire occurrences can be predicted (Sayad et al.; 2019). Basics the concepts of analytics and analysis of data, the problem can be solved using classification or prediction. In Jaafari et al. (2019) the author highlights wildfire occurrence probability prediction will be supportive to find and prevent the future occurrences of wildfire. However, in Sayad et al. (2019) wildfire occurrences are predicted using classification methods involing dead vegetation, surface temperature and weather factors as influencers. And the classification based on this stimulating factors is easy to predict the occurrence of the wildfire (Sayad et al.; 2019). Wildfire occurrence probability prediction model has been developed in Jaafari et al. (2019) which provides accurate probability percentage of future wildfire occurrences based on the influencing factors and past occurrences data. In Jaafari et al. (2019) probability model uses dead vegetation, the temperature of surface, air temperature and weather factors coupled with river resources nearby to wildfire-prone areas. In Jaafari et al. (2019) models probability prediction might lead users to take wrong decisions by getting influenced by near percentage. In Sayad et al. (2019) model classifies the exact status of the zone based on the factors influencing the wildfires at present. In Dacre et al. (2018) all the above said factors which influence occurrences of wildfire were used with additional live vegetation temperature and water content in land, which helps this model to strengthen and gives more accurate probabality prediction with more factors, but surprisingly Sayad et al. (2019) model has higher accuracy among all the other model developed to predict the wildfire in past. Since the model from Sayad et al. (2019) uses a simple classification technique to identify the wildfire class with accuracy of 98.32 in average from Kfold shuffle splitting technique. The model developed in Sayad et al. (2019) might be inaccurate in predicting the wildfire influenced by the weather factors. Model in Jaafari et al. (2019) have included the weather factors while featuring the model, but predicting the probability is not as efficient as the classification. It is evident that weather factors were the most influencing factors causes the wildfire. The author in Sayad et al. (2019) highlights the same future works of the model to be strengthened.

2.2 Selecting an appropriate data source:

Developing an accurate model using data, which depends on the data source completely by means of the quality of data in the data source makes the model better understand of the data to give accurate output Sayad et al. (2019),Jaafari et al. (2019) and Dacre et al. (2018). In this research deals with earth weather and metrological data. For this purpose, only a few data sources serve the best namely Modis, Parasol, Aura, Calipso, Cloud sat, Viirs. Modis and Viirs are the most used data source by past research Dacre et al. (2018). Since they both have user-friendly data extraction and has the best user interface to access the data with organized preprocessing methods Wang et al. (2018). Where Modis is data source hosted and maintained by NASA. VIIRS is also approved by NASA, both of them has its own pros and cons (Sayad et al.; 2019) and (Wang et al.; 2018).

2.2.1 Remote sensing instruments

MODIS and VIIRS both the instruments are capable of NIR (Near Infrared Reflectance) in different scale variations called channels, NIR is a cost-effective way to monitor or understand the land resources or weather factors (Wang et al.; 2018) and (Jaafari et al.; 2019). Modis is developed and maintained by NASA whereas Viirs is developed and maintained by ball aerospace center, which is critically assessed by NASA and approved (Wang et al.; 2018). Both of the instruments are methods state of the art methods used to extract the metrological data and convert them to grayscale image where the data embedded into it, also they provide approved preprocessing methods to get the required data from the image (Wang et al.; 2018),(Jaafari et al.; 2019) and (Sayad et al.; 2019). The important difference between Modis instrument and Viirs is lacking water vapour and CO2 channel in Viirs, since in this research deals with wildfire prediction, which is dependent on the water resources Viirs data source might not the accurate (Wang et al.; 2018). As per the literature survey conducted on both Viirs and Modis instruments, Modis will be suitable for this research.

2.3 Factors influencing wildfire:

The three factors needed for wildfire to ignite and burn the whole forest is referred as wildfire triangle by the fire fighters namely fuel, oxygen and heat source (Sayad et al.; 2019),(Dacre et al.; 2018) and (Dacre et al.; 2018). The components in the wildfire triangle were considered as factors by means of dead vegetation i,e fuel, surface temperature as heat source and finally oxygen as weather factors. These factors were critically assessed and proved that these three factors remain as the pillars for influencing the wildfire occurrences (Kraaij et al.; 2018). In Kraaij et al. (2018) author highlights that dead vegetation is fuel to the fire, the surface temperature is a catalyst to wildfire which might cause ignite the fire from the ground and the climate factors like air temperature, the direction of the winds and soil moisture spreads fire.

2.3.1 Dead Vegetation to wildfire fuel:

Dead vegetation can be an organic waste of dead animals or leftovers of dead species like plants and chemical matters, which resides under the ground this in turn converts as fuel (Wang et al.; 2018). In addition, the fuel is incapable of capturing fire by its own and it needs some catalyst such as surface temperature (Sayad et al.; 2019) and (Wang et al.; 2018). There is a high chance of dead vegetation catching fire in hot surface than wet surface (Sayad et al.; 2019). Organic remains and the chemical waste will remain under ground for years. Same will be changed to fuel or charcoal or some other matter (Collins et al.; 2018). Due to drought or dry weather for prolonged time in a forest converts the green vegetation and dead vegetation as dry bone flammable fuel source, which remains under the ground for years (Collins et al.; 2018). This condition is measured in this research by means of Normalized differential vegetation index (NDVI) from above data source.

2.3.2 Wildfire igniting Source:

When the fuel source is formed all needs is a heat source, that can be caused by lighting or unattended campfire or surface temperature (Poon and Kinoshita; 2018) and (Collins et al.; 2018). The lighting and thunder will occur once in blue moon, so lighting and thunder will not be considered as the primary heat source. Surface temperature is silent source of heat. Due to weather and drought conditions in forest area the surface will increase and humidly on the air decreases this condition hikes the surface temperature to extinct (Collins et al.; 2018). In this situation the fuel underground from dead vegetation get ignited as getting contact with hot surface (Kraaij et al.; 2018) and (Collins et al.; 2018). As per the survey heat source plays the important role, this factor will be measured by Land surface temperature from the MODIS data source.

2.4 Role of Weather factors in wildfire:

It is clear that role of the weather in wildfire occurrences is vital namely converting the dead vegetation to fuel and spreading the wildfire as oxygen source (Collins et al.; 2018) and (Kraaij et al.; 2018). In general weather features depends on many substances like air temperature, wind direction, Soil moisture (Rodrigues et al.; 2019). Air temperature has influence in occurrences of wildfires, however the same is interacted with land surface and turns as a heat source (Collins et al.; 2018) and (Kraaij et al.; 2018). Featuring same substances in different variations will not suitable, so air temperature will not be featured in this research model. Wind directions controls the spreading of the fire, but this research deals with future fire prediction based on classification of the classes (Rodrigues et al.; 2019). Hence adding the wind direction won‘t help the model to predict the state of fire. Soil moisture can be measured in different levels from ground, since this research deals with ground fires below ground levels would be more suitable for estimating the soil moisture and add as additional feature to model by means of checking the improvement in model accuracy (Zhang et al.; 2017) and (Ghahremanloo et al.; 2018). As per literature review conducted for weather factors, soil moisture will be suitable for featuring the model as an additional attribute.

2.4.1 Estimating soil moisture:

Soil moisture is one of the key factors while monitoring and managing the wildfire or the natural resources (Ghahremanloo et al.; 2018). In Ghahremanloo et al. (2018) author uses a specifically build system to measure the estimate the soil moisture. In Zhang et al. (2017) MODIS data source and some additional data source for identifying the terrain is used so that accurate values of soil moisture are extracted. Soil moisture is related to the perpendicular drought index (PDI) (Zhang et al.; 2017). This PDI is extracted from the MODIS then combining both LST and PDI soil moisture is estimated for the exact date and region. Soil moisture can be extracted in different levels namely surface soil moisture and sub surface soil moisture (Vinodkumar and Dharssi; 2019). In Vinodkumar and Dharssi (2019) author uses different levels of soil moisture extracted from old embrical instrument specifically for Australia to map the fire zone. Soil moisture is estimated using SMOS (Soil Moisture Ocean Salinity) method with level 2 observation by surface soil moisture from ground to 5mm below ground and sub surface soil moisture is below 5mm from ground with different min and max scales respectively. As per the literature review the Soil moisture with above variations will be used in this research.

3 Methodology

KDD process model is followed throughout this research from selecting data source to end evaluation as shown in the figure below.

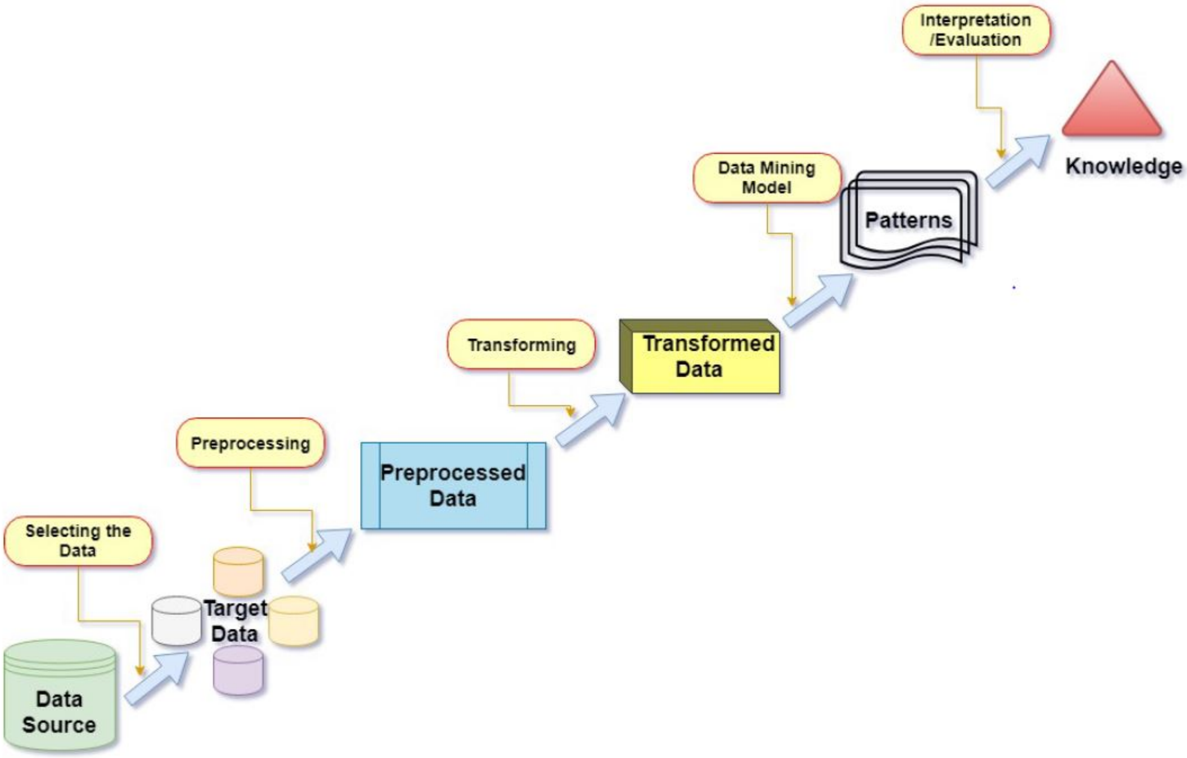


Figure 1: KDD

3.1 Selecting the Data:

As per the literature conducted above 2.2 MODIS has been used to extract NDVI, LST, Burned Area. Every attribute is taken from the respective instrument as shown in the table below 3.1.

Data Instruments			
Attribute name	Satellite	Instrument name	Data captured data interval
NDVI	MODIS	MOD13Q1	16
LST	MODIS	MOD11A2	8
Burned Area	MODIS	MOD14A2	8

As all the above instruments store data in two different formats namely HDF metadata file and ECS metadata file. Where the HDF stands for Hierarchal Data Format and contains valuable information like raster value, binary data. HDF file will be utilized in this research. Data is extracted only for 44 location in British Columbia which is declared as the wildfire occurred zone on the specific year as sample shown in the table 3.2.3.

3.1.1 Estimation of NDVI:

HDF metadata files extracted from the earth explorer or earth data search for a particular timeframe. This is converted to GeoTiff file and the GeoTiff file is processed based on the below mentioned formula 1 to get the estimation of NDVI between the range of -1 to 1, were data towards -1 means vegetation is poor in that location and data towards +1 means is vegetation is good in that region.

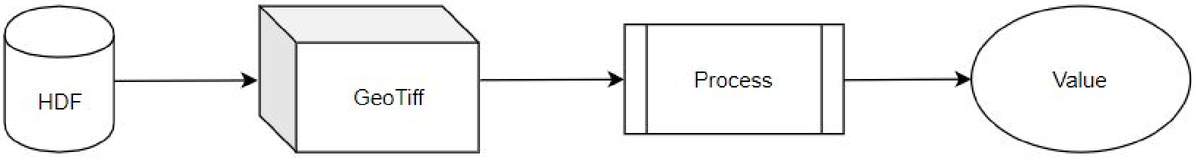


Figure 2: HDF Processing

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)} \tag{1}$$

All the above-said process needs a massive platform to process the HDF files to tiffs which is not feasible, so the same process is done by the API based on requests [3]. By selection of the respective product from the table and period coupled with the latitude and longitude details will make the work simple by sub setting the data for the request.

3.1.2 Estimation of LST and Burned Areas:

Estimating LST and Burned area is same as the NDVI. Instead of formula used in NDVI, LST and Burned area value is calculated by using binary digit decryption as mentioned in LPDAAC website. As this processing requires a massive platform to process the data. So, by using an API as mentioned above, LST and Burned area values have been extracted in recommended units.

[3] https://modis.ornl.gov/cgi-bin/MODIS/global/subset.pl

3.1.3 Estimation of Soil moisture:

Soil moisture can be measured in different levels from the ground as discussed earlier in the literature survey 3.1.3. Based on that soil moisture is extracted from google earth engine in two variations namely surface soil moisture and sub-surface soil moisture. Based on the latitude and longitude by year both variation of saturated soil moisture percentage will be extracted [4].

3.2 Data Pre-processing and Transforming:

Data pre-processing is divided into three stages namely.

1. Data Merging
2. Data Cleaning
3. Data Labelling

After the extraction of the raw data very next step is data pre-processing. In this research data pre-processing is done in R, As it is easy to handle missing data and merging five different raw data files.

3.2.1 Data Merging:

Data merging is a challenging part of in pre-processing, since the raw data is generated based on the spatial resolution of 1 km by each instrument, which means, for example, every 20 kilometres square area from latitude and longitude of requested data. There will be 40 values for each area. The main challenge is merging the exact cell data to another instrument extracted cell data, so removal of the failed pixel value is not possible in this step. Then comes the next challenge LST and Burned Area have similar dates of the 8-day interval of data per month for every location, but NDVI data has a 16-day interval of data in a month. So LST and Burned area data have been scaled down by removing the dates which are not in NDVI dates. As the common dates are shown in the table below 3.2.1.

Common Dates across all attributes	
Month	Days in common
January	1,9,17,25
February	2,10,18,26
March	6,14,22,30
April	7,23
May	1,9,17,25
June	2,10,18,26
July	12,28
August	13,29
September	6,14,22,30
October	16
November	1,9,17,25
December	3,9

[4] https://developers.google.com/earth-engine/datasets/catalog/NASA_USDA_HSL_soil_moisture

Above dates was common across years, since now merging the NDVI, LST, burned area is based on Latitude, longitude, year, month, the day is done. Next is merging the variations of soil moisture to the above-merged dataset based on same Latitude, longitude, year, month, day. Since both soil moisture variation has 3-day interval data of saturated soil moisture. As per above common dates both soil moisture variations will be scaled down and merged with the above dataset.

3.2.2 Data Cleaning:

The next step after data merging is data cleaning. Cell with “F” value is because of instrument failure or cloud cover. Same will be removed from the dataset instead of imputing values from another cell(Wang et al.; 2018). Every pixel is independent cell from the whole area, imputing the mean or median value from another cell value won’t get the best model accuracy (Sayad et al.; 2019). After removing the failed pixel value redundant values has been removed by comparing all factors in that dataset (Yang et al.; 2019).

3.2.3 Data Labelling:

As the dataset used in this research is not labelled by the data source, based on the past wildfire history recorded in [5] dataset is labelled as using the table below 3.2.3.

Sample Data of wildfire events

Wildfire type	Year	Month	Date	Place	Latitude	Longitude
Lighting Fire	2010	July	28th - 1st Aug	Cariboo	54.02681	-123.9478
Summer Fire	2012	Octomber	1st - 31st	west of Clinton	50.8131	-121.3242
Summer Fire	2012	Octomber	1st - 31st	Peachland	49.7666	-119.75
Summer Fire	2012	Octomber	1st - 31st	Fort Nelson	58.8062	-122.6939
Lighting Fire	2013	May	1st - 31st	Ashcroft	50.7212	-121.2835
Lighting Fire	2013	May	1st - 31st	Tweedsmuir North Provincial Park	49.2813	-123.1227
Climate	2014	July	15th - 31st Aug	Quesnel	52.9794	-122.4936
Climate	2014	July	15th - 31st Aug	Bull Canyon	49.2609	-123.2471

The above table has different kinds of wildfires, in this research only climate fire and summer fire events were used. Since this research deals with surface fires which occurs based on the climate and land surface temperature. As per the above table, the dataset

[5] https://tinyurl.com/yy5ogo5e

is labelled based on the year, month, day, latitude and longitude as fire (1) and no fire class (0) respectively .

3.3 Data mining model:

This section explains about exploratory data analysis done on this dataset to the application of the data mining alogorithm.

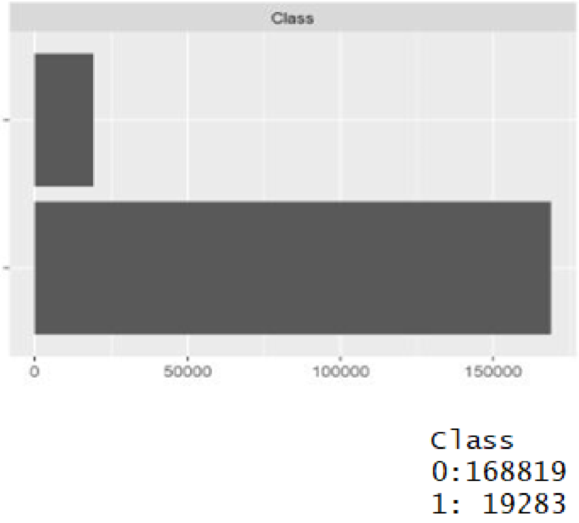
3.3.1 Data Specification:

The Wildfire dataset contains 188102 rows with 11 columns where year, month, day, latitude and longitude are used to merge the data as those columns will not be used in model development.

Data Specifications			
SlNo	Code	Description	Domain
1	year	year	2012 - 2018
2	Month	Month	12 Months in year
3	Day	Day	Day
4	Latitude	Latitude	Latitude
5	Longitude	Longitude	Longitude
6	NDVI	Normalizsed Diffrential Vegetation Index	-1 to +1
7	LST	Land Surface Temperature	150 Kelvin to 258 Kelvin
8	Burned Area	Burned Area	5 to 9
9	SM	Surface soil moisture	0* to 25.39* mm
10	SUSM	Subsurface soil moisture	0* to 274.6* mm
11	Class	Class	Fire = 1, No fire = 0

The above table 3.3.1 depicts the NDVI values were between -1 to +1 interval, LST values are between 150 to 258 kelvin interval, surface soil moisture values are in between 0 to 25.39 and subsurface soil moisture values were estimated between 0 to 274.6 and burned area is a categorical value each category is categorised as below table.

Data Specifications	
Value	Description
5	Non-fire land pixel
6	cloud (land or water)
7	Fire (low confidence, land or water)
8	Fire (nominal confidence, land or water)
9	Fire (high confidence, land or water)



```
Class
0:168819
1: 19283
```

Figure 3: Target Variable

As the above bar graph 3.3.1 depicts class “0” i,e no fire class has 168819 counts of records, whereas class “1” has 19283 row counts. From the above information, it is very clear that there is a class imbalance.

As the pre-processing of raw data is completed next big stage is applying the appropriate data mining algorithm into data. Since the dataset used in this research is labelled, the supervised machine learning algorithm will be applied. As seen above in the literature survey model from Sayad et al. (2019) is proven to have higher accuracy than other research models on classification, so the same model has been applied in this research with a newly created dataset.

3.4 Model Evaluation:

Post developing the model its capability on the classification of new data can be measured through prediction accuracy, precision, recall and f1 score. However, all the above mentioned are just scores still as part of understanding the real efficiency of the model confusion matrix will be used to identify the prediction capability of the model on the testing data.

4 Design Specification

In this research, two questions will be answered based on the results from the respective models.

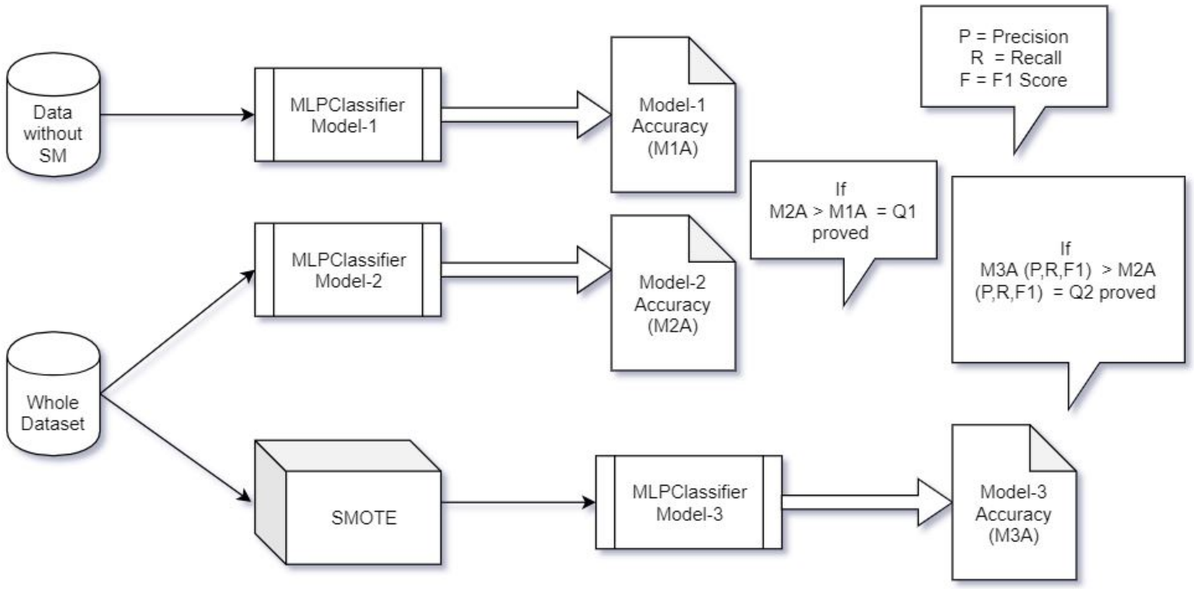


Figure 4: Model Architecture

Q1 Will the inclusion of soil moisture attribute in the state-of-art model for wildfire prediction provide a statistically significant improvement ?

Q2 Will SMOTE analysis provides a statistically significant improvement on Precision, Recall, F1 score of the model?

As per the above questions, the research has been carried out with three models to answer the questions shown in the architecture diagram Above.

According to above figure 4, MLPClassifer model-1 is developed to get a base accuracy, which has been featured with NDVI, LST, Burned area as predictor variable and class as target variable. The MLPClassifer model-1 is named as model without soil moisture attribute. The whole dataset is applied to MLPClassifer model-2 with both variations of soil moisture in it, this model is names has model with soil moisture attributes. The accuracy of model-1 is campared with the prediction accuracy of model-2 to check whether is there a significant improvement after the inclusion of soil moisture attributes. As the dataset has class imbalance as seen in data specification stage 3.3.1. Same dataset is subjected to smote analysis to create create synthetic sample of minority class. The data after smote analysis will be featured to MLPClassifer model-3 to check whether the there is significant improvement in the accuracy and the metrics frm the other two models. This model named as model after SMOTE analysis.

4.1 MLPClassifier:

The model in Sayad et al. (2019) uses MLPClassifier to get the higher accuracy from other models. This MLPClassifier is from sklearn package with only one layer of 100 units this uses backpropagation to remember errors from the past. This model is validated using KFold stratified shuffle splitting with 10 folds. This particular package is being to handle the class imbalance accuracy automatically instead of keras (Kanin et al.; 2019).

```python
kfold = StratifiedShuffleSplit(n_splits=10, test_size=0.15, random_state=12)
training_accuracy = []
testing_accuracy = []
epochs=10
for train, test in kfold.split(X_train, y_train):
  clf = MLPClassifier(solver='sgd',learning_rate='adaptive',momentum=0.9, activation='relu',alpha=5, batch_size='auto',verbose=True, n_iter_no_change = epochs)
  clf.fit(X_train,y_train)
  model = clf
  #training accuracy
  y_tr_pred = clf.predict(X_train)
  results_tr = cross_val_score(model, X_train ,y_tr_pred, cv = kfold, verbose=1)
  training_accuracy.append(results_tr.max()*100.0)
  #testing Accuracy
  y_te_pred = clf.predict(X_test)
  results = cross_val_score(model, X_test ,y_te_pred, cv = kfold, verbose=1)
  testing_accuracy.append(results.max()*100.0)
  print("Training Accuracy (Shuffle Split) : %.3f%% (%.3f%%)" % (results_tr.max()*100.0, results_tr.std()*100.0))
  print("Prediction Accuracy (Shuffle Split) : %.3f%% (%.3f%%)" % (results.max()*100.0, results.std()*100.0))
```

Listing 1: MLPClassifier

The above listing shows the hyperparameters and MLP method used in this research. Optimizer used is stochastic gradient descent, where learning rate is adaptive. To allow the model by take the best learning rate exhibit the higher accuracy.

5 Implementation

This research aims to explore two questions discussed in design specifications 4. This section will explain the implementation attempt to prove the questions.

5.1 Initial data extraction and aggregation:

Processing the MODIS HDF files require massive environment and meteorological knowledge, which is time consuming. For doing this requirement there was a site which is freely available for non-metrological background analytics [6]. Using this tool NDVI, LST, burned area values for 44 severe wildfire zones were captured from 2012 to 2018 year. For extraction of soil moisture google earth engine has been, which gives a saturated percentage of soil moisture in two variations namely surface soil moisture of 5mm depth below ground and subsurface soil moisture of 10mm depth below surface [7]. After extraction of raw data cleaning and merging was a very challenging task, since each attribute was extracted from respective instruments separately. All the attributes were taken down to common date scale of availability and merged based on latitude, longitude, day, year and month.

[6] https://modis.ornl.gov/cgi-bin/MODIS/global/subset.pl

[7] https://developers.google.com/earth-engine/datasets/catalog/NASA_USDA_HSL_soil_moisture

5.2 Data cleaning and transformation:

After the data is merged as said above failed pixel values and redundant values has been removed across all attributes using R. Summer fire and climate fire data has been filtered and labelled. Based on wildfire events accounted in British Columbia site [8].

5.3 Building the model:

MLPClassifier is being used in this classification. To answer the research questions three models were built as explained below.

1. Model without Soil moisture
2. Model with soil moisture
3. Model in class balanced dataset

5.4 Model without Soil moisture attributes:

This model developed to set a base benchmark to all the other models by using NDVI, LST, burned area attributes as done in (Sayad et al.; 2019). As this model will be evaluated by kfold stratified shuffle splitting technique.

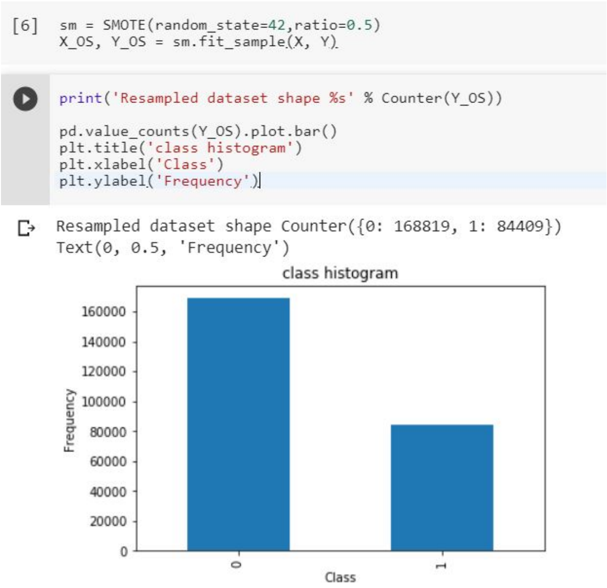
5.5 Model with Soil moisture attributes:

Both variations of soil moisture namely surface soil moisture and sub surface soil moisturer have been included in the model with other attributes.If the model exhibits a statistically significant improvement with the addition of soil moisture attributes (Q1) will be proved. still the prediction of the model two is skewed, since the model uses class imbalance dataset. As the Wildfire is rarely occurring event under sampling of majority class might help or over sampling the minority class will be helpfull using smote method.

5.6 Model with class balanced data:

smote technique is applied to the class imbalanced data to create synthetic samples of minority class to make the dataset balanced (Raghuwanshi and Shukla; 2019).
The below output picture depicts new synthetic samples of 60000 new data points were created based on the minority class data points to balance the dataset. The ratio 0.5 means minority class has been over-sampled up to 50 percent of the majority class. After applying the smote, data is applied and trained in MLPClassifier to get a balanced prediction accuracy.

[8] https://tinyurl.com/yy5ogo5e



```
[6]  sm = SMOTE(random_state=42,ratio=0.5)
     X_OS, Y_OS = sm.fit_sample(X, Y)
```

```
print('Resampled dataset shape %s' % Counter(Y_OS))

pd.value_counts(Y_OS).plot.bar()
plt.title('class histogram')
plt.xlabel('Class')
plt.ylabel('Frequency')
```

```
Resampled dataset shape Counter({0: 168819, 1: 84409})
Text(0, 0.5, 'Frequency')
```

Figure 5: SMOTE

5.7 Implementation Platform:

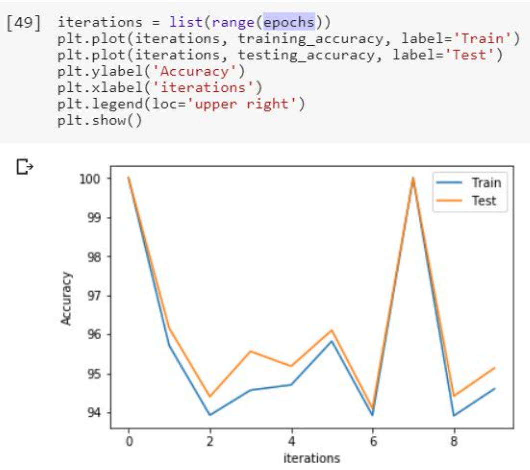
After choosing an appropriate algorithm and running in a platform of random choice won't get the higher accuracy, getting higher accuracy of the model is highly dependent on the underlying platform (Sayad et al.; 2019),(Dacre et al.; 2018). Cloud vendors like amazon, OpenStack provides a platform where the supporting libraries and softwares have to installed which consumes a lot of time, so to overcome that google is providing a readily available environment called google Collaboratory in short google colab for Machine learning and data analytics (Carneiro et al.; 2018). Where all required packages and library were installed and kept ready to start building the model in python. In this research colab is being used has to a platform for building the models, where preprocessing is done in R.

6 Evaluation

Three models have been developed using MLPClassifier from sklearn package. KFold stratified shuffle split method with 10 folds was used to validate all the models. Each fold accuracy has been averaged to get the model accuracy. Alternatively, model classification efficiently is evaluated using the confusion matrix with precision, recall, f1 score.

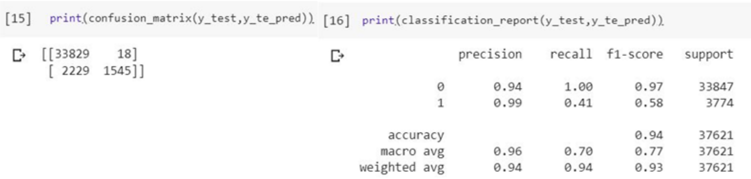
6.1 Model without Soil moisture attributes:

The below graph 6.1, depicts testing and training accuracy in each fold, however in few folds model tends to overfit due to class imbalance but overall its exhibits 96 % of testing accuracy in average respectively.



```
[49] iterations = list(range(epochs))
     plt.plot(iterations, training_accuracy, label='Train')
     plt.plot(iterations, testing_accuracy, label='Test')
     plt.ylabel('Accuracy')
     plt.xlabel('iterations')
     plt.legend(loc='upper right')
     plt.show()
```

Figure 6: Accuracy For Model-1



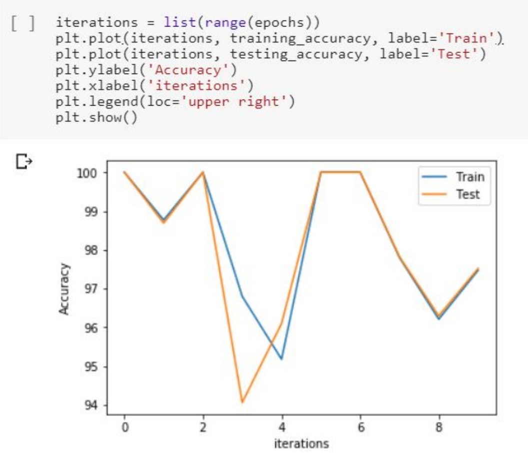
```
[15] print(confusion_matrix(y_test,y_te_pred))

[[33829    18]
 [ 2229  1545]]
```

```
[16] print(classification_report(y_test,y_te_pred))

              precision    recall  f1-score   support

           0       0.94      1.00      0.97     33847
           1       0.99      0.41      0.58      3774

    accuracy                           0.94     37621
   macro avg       0.96      0.70      0.77     37621
weighted avg       0.94      0.94      0.93     37621
```

Figure 7: Confusion matrix and classification report from Model -1

As from the above classification report 6.1, image model exhibited 99 % of precision, 41 % on recall and 58 % in f1 score. As the above confusion matrix output depicts True Negatives of 33829 is classified correctly, there are 18 False Positives with 2229 False Negatives and 1545 True Positives.

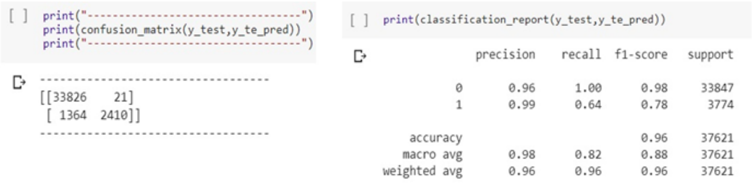
6.2 Model with Soil moisture attributes:



```
[ ] iterations = list(range(epochs))
    plt.plot(iterations, training_accuracy, label='Train')
    plt.plot(iterations, testing_accuracy, label='Test')
    plt.ylabel('Accuracy')
    plt.xlabel('iterations')
    plt.legend(loc='upper right')
    plt.show()
```

Figure 8: Model-2 Accuracy

As the above graph 6.2, depicts testing and training accuracy in each fold, however in few folds model tends to over-fit due to class imbalance but overall its exhibits 98 % of testing accuracy in average respectively.



```
[ ] print("--------------------------------")
    print(confusion_matrix(y_test,y_te_pred))
    print("--------------------------------")

--------------------------------
[[33826    21]
 [ 1364  2410]]
--------------------------------
```

```
[ ] print(classification_report(y_test,y_te_pred))

              precision    recall  f1-score   support

           0       0.96      1.00      0.98     33847
           1       0.99      0.64      0.78      3774

    accuracy                           0.96     37621
   macro avg       0.98      0.82      0.88     37621
weighted avg       0.96      0.96      0.96     37621
```

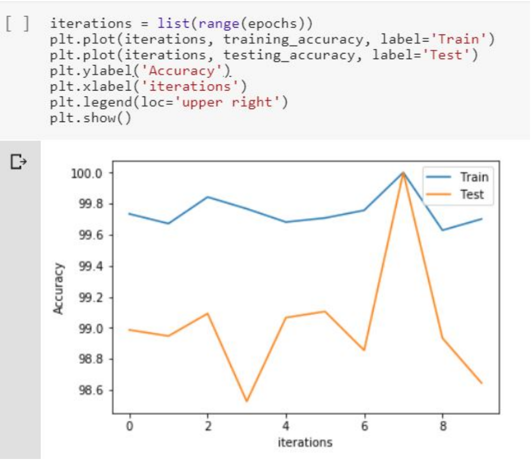
Figure 9: Confusion matrix and classification report from Model -2

From the above classification report 6.2, image model exhibited 99 % of precision, 64 % on recall and 78 % in f1 score.

As the above confusion matrix output depicts True Negatives of 33826 is classified correctly, there are 21 False Positives with 1364 False Negatives and 2410 True Positives.

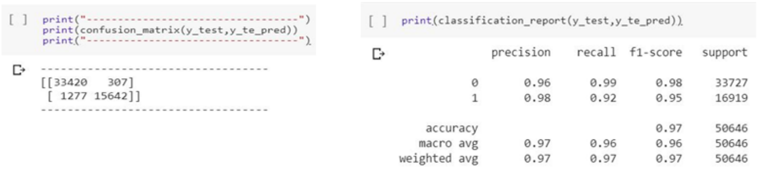
6.3 Model after SMOTE analysis:

The below graph 6.3, depicts testing and training accuracy in each fold, however in few folds model tends to over-fit due to class imbalance but overall its exhibits 99 % of testing accuracy in average respectively.



```
[ ] iterations = list(range(epochs))
    plt.plot(iterations, training_accuracy, label='Train')
    plt.plot(iterations, testing_accuracy, label='Test')
    plt.ylabel('Accuracy')
    plt.xlabel('iterations')
    plt.legend(loc='upper right')
    plt.show()
```

Figure 10: Model-3 Accuracy



```
[ ] print("--------------------------------")
    print(confusion_matrix(y_test,y_te_pred))
    print("--------------------------------")

--------------------------------
[[33420   307]
 [ 1277 15642]]
--------------------------------
```

```
[ ] print(classification_report(y_test,y_te_pred))

              precision    recall  f1-score   support

           0       0.96      0.99      0.98     33727
           1       0.98      0.92      0.95     16919

    accuracy                           0.97     50646
   macro avg       0.97      0.96      0.96     50646
weighted avg       0.97      0.97      0.97     50646
```

Figure 11: Confusion matrix and classification report from Model-3

From the above classification report 6.3, model exhibited 98 % of precision, 92 % on recall and 95 % in f1 score. As the above confusion matrix output depicts True Negatives of 334120 is classified correctly, there are 307 False Positives with 1277 False Negatives and 15642 True Positives.

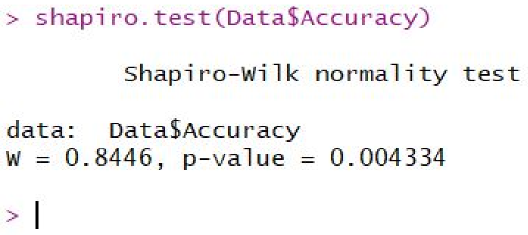
6.4 Discussion on the results:

As the results are explained above by comparing the model 1 and 2, it is evident that there is statistically significant improvement of prediction from model 1 to model 2, but classification report output of model 2 is very low. This is due to class imbalance as shown in the figure 3.3.1. However, this research aims to validate the improvement in accuracy after the inclusion of soil moisture into the state-of-the-art model of wildfire prediction. The model is not trained enough with fire class samples. So, the prediction accuracy of fire is too low from the classification report. To overcome this challenge SMOTE has been used into this dataset to create synthetic samples of fire class datapoints. Post SMOTE process as shown image 5.6, the over-sampled data is treated to same MLPClassifier based model. From the classification report scores and confusion matrix of model 3, has a significant improvement in prediction accuracy and classification metrics scores and the research questions proposed is proved from the results.To prove the accuracy of model 1 and 2 accuracy are significantly different a statistical test has to be done to prove the accuracy is significantly different or improved. As this data is found to be not normally distributed, applying parametric test is not feasible. To find the proof for accuracy is significantly different. Mann-Whitney U test is applied.

6.4.1 Mann-Whitney U test:

Before subjecting the data to test, data needs to satisfy below assumptions.

1. Dependent variable is continuous i,e accuracy score.
2. Independent variable is categorical i,e the Model (1 or 2).
3. Independence of observations the accuracy score is collected from two different model with different attributes.
4. Not normally distributed, as it is clear from Shapiro-Wilk test the data is not normally distributed as shown in the output below.



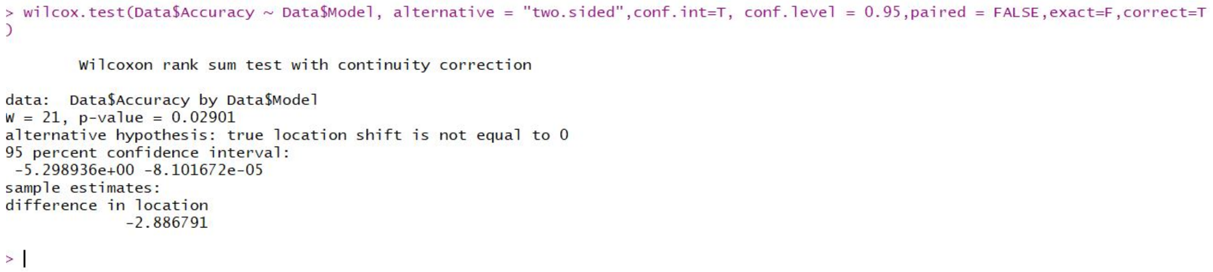
```
> shapiro.test(Data$Accuracy)

        Shapiro-Wilk normality test

data:  Data$Accuracy
W = 0.8446, p-value = 0.004334

> |
```

Figure 12: Shapiro-Wilk test

As the all the assumptions were meet data has been applied to the Mann- Whitney u test as the output shown below.

Ho: There is no difference in the mean accuracy score of model 1 and model 2

Ha: There is difference in the mean accuracy score of model 1 and model 2



```
> wilcox.test(Data$Accuracy ~ Data$Model, alternative = "two.sided",conf.int=T, conf.level = 0.95,paired = FALSE,exact=F,correct=T
)

        Wilcoxon rank sum test with continuity correction

data:  Data$Accuracy by Data$Model
W = 21, p-value = 0.02901
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -5.298936e+00 -8.101672e-05
sample estimates:
difference in location 
             -2.886791 

> |
```

Figure 13: Mann-Whitney U Test

As the p-value is less than 0.05 we have enough proof to reject null hypothesis, Thus the mean difference of model 1 and model 2 is significantly different.

7 Conclusion and Future Work

The research was aimed to understand is there a statistically significant improvement in prediction accuracy by including the soil moisture a weather attribute towards the state-art-model of wildfire prediction.In a newly created dataset, there is a class imbalance. Thus from the results from the model 3 clearly proves the above objectives were meet as the model with balanced dataset gave 99 % of accuracy with 98 % precision, 92 % recall, 95 % F1 score. This score is a major improvement from the scores of model 2, where data has a class imbalance. From classification report output of model 2, it's crystal clear that the model is not trained with the fire class data, however, epochs can also be increased to train the model, which consumes more time and resource. Hence alternatively this SMOTE is used to get higher accuracy with less time consumption. As the main objectives of this research have been meet, still from the confusion matrix of model 3 has a notable count of False Negatives. Which depicts that model needs more training epochs, due time constraint in this research more training epochs were not done As part of future work in this research, training epochs can be increased to get less False positives and the data is time-series based, LSTM (Long Short-Term Memory) also can be applied to compare the classification metrics and accuracy, which has more memory than MLPClassifier used in this research.

References

Boisram, G., Thompson, S. and Stephens, S. (2018). Hydrologic responses to restored wildfire regimes revealed by soil moisture-vegetation relationships, *Advances in Water Resources* **112**: 124–146.

Carneiro, T., Medeiros Da Nobrega, R. V., Nepomuceno, T., Bian, G.-B., De Albuquerque, V. H. C. and Filho, P. P. R. (2018). Performance analysis of google colaboratory as a tool for accelerating deep learning applications, *IEEE Access* **6**: 61677–61685.

Collins, L., Griffioen, P., Newell, G. and Mellor, A. (2018). The utility of random forests for wildfire severity mapping, *Remote Sensing of Environment* **216**: 374–384.

Dacre, H. F., Crawford, B. R., Charlton-Perez, A. J., Lopez-Saldana, G., Griffiths, G. H. and Veloso, J. V. (2018). Chilean wildfires: Probabilistic prediction, emergency response, and public communication, *Bulletin of the American Meteorological Society* **99**(11): 2259–2274.

Ghahremanloo, M., Mobasheri, M. R. and Amani, M. (2018). Soil moisture estimation using land surface temperature and soil temperature at 5 cm depth, **40**(1): 104–117.

Jaafari, A., Zenner, E. K., Panahi, M. and Shahabi, H. (2019). Hybrid artificial intelligence models based on a neuro-fuzzy system and metaheuristic optimization algorithms for spatial prediction of wildfire probability, *Agricultural and Forest Meteorology* **266-267**: 198–207.

Kanin, E., Osiptsov, A., Vainshtein, A. and Burnaev, E. (2019). A predictive model for steady-state multiphase pipe flow: Machine learning on lab data, *Journal of Petroleum Science and Engineering* **180**: 727–746.

Kraaij, T., Baard, J. A., Arndt, J., Vhengani, L. and van Wilgen, B. W. (2018). An assessment of climate, weather, and fuel factors influencing a large, destructive wildfire in the knysna region, south africa, *Fire Ecology* **14**(2).

Poon, P. K. and Kinoshita, A. M. (2018). Spatial and temporal evapotranspiration trends after wildfire in semi-arid landscapes, *Journal of Hydrology* **559**: 71–83.

Raghuwanshi, B. S. and Shukla, S. (2019). Smote based class-specific extreme learning machine for imbalanced learning, *Knowledge-Based Systems* .

Rodrigues, M., Gonzlez-Hidalgo, J. C., Pea-Angulo, D. and Jimnez-Ruano, A. (2019). Identifying wildfire-prone atmospheric circulation weather types on mainland spain, *Agricultural and Forest Meteorology* **264**: 92–103.

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

Vinodkumar and Dharssi, I. (2019). Evaluation and calibration of a high-resolution soil moisture product for wildfire prediction and management, *Agricultural and Forest Meteorology* **264**: 27–39.

Wang, J., Liu, C., Min, M., Hu, X., Lu, Q. and Husi, L. (2018). Effects and applications of satellite radiometer 2.25-m channel on cloud property retrievals, *IEEE Transactions on Geoscience and Remote Sensing* **56**(9): 5207–5216.

Yang, G., Sun, W., Shen, H., Meng, X. and Li, J. (2019). An integrated method for reconstructing daily modis land surface temperature data, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **12**(3): 1026–1040.

Zhang, D., Meng, L., Qu, J. J., Zhang, W. and Wang, L. (2017). Estimation of surface soil moisture in cornfields using a modified modis-based index and considering corn growth stages, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **10**(12): 5618–5631.