2020

A machine learning algorithm for predicting wild fire occurrence.

Otieno, Jack Odunga
Faculty of information technology
Strathmore University

A Machine Learning Algorithm for Predicting Wild Fire Occurrence

Odunga, Jack Otieno

Student Number: - 101086

Master of Science in Information Technology

A Machine Learning Algorithm for Predicting Wild Fire Occurrence

Odunga, Jack Otieno

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Information Technology (Msc.IT) at Strathmore University

Faculty of Information Technology

Strathmore University

Nairobi, Kenya

April 2020

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Odunga, Jack Otieno

……………………………

April 2020

Approval

The thesis of Odunga, Jack Otieno was reviewed and approved by the following:

Dr. Vitalis Gavole Ozianyi

Senior Lecturer, Faculty of Information Technology,

Strathmore University

Dr. Joseph Orero

Dean, Faculty of Information Technology

Strathmore University

Dr. Bernard Shibwabo

Director of Graduate Studies

Strathmore University

Abstract

A wild fire is an unplanned fire that burns in a natural area such as forest and grassland. Indeed, wild fires are destroying parts of the world's forest coverage, affecting land and killing wild life. Majority of wild fires are caused by human activities and weather conditions. Despite various forest management authorities using numerous methods to detect and suppress fire, there are several wild fires reported around the globe annually. In Kenya for instance, Kenya Forest Services ensures fire detection and detention through the use of ground patrols and fixed stations (fire towers). They also use radio systems, vehicles, motorcycles and even bicycles. These techniques are not working effectively due to inadequate staff and resources to cover the available forest coverage. The research proposed the development of supervised machine learning model to predict wild fires using existing data that was collected from credible climatological sources that included both meteorological sources as well as wildfire databases focusing on content dated from 2000 to 2020. The study utilized data sets from multiple sources including National Fire Danger Rating System (NFDRS), Canada National Fire Database (CNFDB), University of California machine learning repository, and scientifically verified Internet sources. The methodology involved collection of relevant data sets, cleaning and preparing the data, training the models, model testing and validation. The climatological factors were used, as input values and Artificial Neural Network (ANN) implemented to establish prediction model. The model was developed through rapid application development (RAD) methodology. Upon completion it was deployed on a web environment to be used by various stakeholders in monitoring and predicting wild fires by giving a binary output of a yes or no on the likelihood of wildfire occurring. Artificial Neural Network Model was trained and validated using 80% and 20% of the set features respectively. The model gave performance accuracy of 82.69 per cent.

Keywords: *Wild fires, Forest, Forest Management Authorities*

Table of Contents

List of Tables

List of Figures

List of Abbreviations/Acronyms

AFFIRMS	–	Administrative and Forest Fire Information Retrieval and Management
ANN	–	Artificial Neural Networks
CBFR	–	Context-Based Fire Risk
CNFDB	–	Canada National Fire Database
DF	–	Drought Factor
FAO	–	Food and Agriculture Organization
FFDI	–	Forest Fire Danger Index
IC	–	Ignition Component
KNN	–	K-Nearest Neighbours
NFDRS	–	National Fire Danger Rating System
RAD	–	Rapid Application Development
SVM	–	Support Vector Machine
WWF	–	World Wide Fund for Nature

Acknowledgements

This is to our Almighty God for directing me in the right path and making it possible. I feel your guidance in every stage of this thesis. Thank you for your endless blessings.

I would like to thank my supervisor, Dr. Vitalis Gavole Ozianyi for the guidance and support. Thank you for your contribution of time and ideas to make my work fruitful and exciting.

To my family, thank you for the enduring support during this course. My wife, Maureen Otieno, thank you for encouraging me in all of my pursuits and inspiring me to follow my dreams.

I am also indebted to my friend and MSIT 2019 classmate Billy Owire, not only for all his help and useful suggestions but also for being there to listen when I needed an ear.

Dedication

To my wife Maureen Otieno, sons Brian & Adrian Otieno, and daughter Brianna Otieno, thank you for your support, prayers, understanding and perseverance.

Chapter 1: Introduction

1.1 Background

Forests are natural resources that are critical components of the Earths ecosystem but are constantly threatened by the wildfires (Liang, Zhang & Wang, 2019). Wild fires cause harm to wildlife, compromise air quality, and threaten the safety of world communities. Wild fires produce smoke that can drastically reduce photosynthetic activity (Davies & Unam, 1999) and can be damaging to the health of humans and animals. The worst wild fires recorded in the recent history have ravaged the world forests due to human activities; persistent heat and drought that makes it easy for fire to start and spread (Krebs, Pezzatti, Mazzoleni, Talbot & Conedera, 2010). Most of the high productive forests including both indigenous and plantations are located in the relatively high fire prone areas (Reduce Wildfire Risk – Washington Forest Protection Association n.d.). Dale et al. (2001) observed that the wild fire effects on forests include, acceleration of nutrient cycling, mortality of individual trees, shifts in successional direction, induced seed germination, loss of soil seed bank, increased landscape heterogeneity, changes in surface-soil organic layers and underground plant root and reproductive tissues, and volatilization of soil nutrients.

According to Chas-Amil, Touza, Prestemon and MacClean (2015) there is need to understand the main reasons aiding wild fire ignitions in order to factor effective fire prevention policies. Dale et al. (2001) noted that the regularity, scope, strength, seasonality, and nature of fires depend on weather and climate in addition to forest structure and composition. Bonsor (2001) outlined three main components needed for ignition and combustion to occur. The first component is fuel, which its size, shape, arrangement and moisture content are the basic characteristics that decide how it affects a fire. The second component is weather through which temperature, wind and moisture are the three ingredients that affect wildfires. The final component is topography that can either aid or hinder wildfire progression. The conditions that favour wildfire occurrence can be estimated through systematic analysis of the values of climatological variables, human activities and forest composition (Toledo-Castro, et al., 2018)

With the effect of global warming, wildfire risks are expected to rise; the human population is also rising and more people are encroaching the forestlands placing more settlements on danger of wildfire (Wolters, 2019). This therefore, calls for the need to adopt a more proactive approach to wildfire management so as to realise cost-effective value compared to live-fire fighting and disaster recovery from post-fire events. There are technical solutions that already use artificial intelligence and machine learning to analyse datasets for prediction of wildfires.

In their study, "Exploiting Poisson additivity to predict fire frequency from maps of fire weather and land cover in boreal forests of Quebec, Canada", Marchal, Cumming and Mclntire, (2016) gathered all wildfire data that had occurred from 2000 to 2010, and upon analysing them established a Poisson regression model to predict the occurrence of forest wildfires in the respective study areas. Liang, Zhang and Wang, (2019) collected wildfire and meteorological data that had been recorded between 1990 and 2018 inclusive from Canada National Fire Database (CNFDB), which they divided, into training and testing sets. Meteorological factors were used as input values, a backpropagation neural network (BPNN), a recurrent neural network (RNN), and long short-term memory (LSTM) were applied to institute wild fire prediction models (Liang, Zhang & Wang, 2019).

Similarly, Plucinski, Mccaw, Gould and Wotton (2014) developed a model using negative binomial regression, collecting data spanning three years and were able to predict operational tools for the day-to-day fire planning. While there exists a number of solutions to predict wildfire occurrence, they do portray a number of shortcomings such as the use of partial data that may result to inaccurate predictions. Some of the solutions are also complex and not easy to understand and use.

This study proposed an easy to use prediction model that proportionately accounted for climatological factors and human activities in the accurate prediction of wildfire occurrence thus allowing for strategic deployment of resources in preventing wildfires from starting and reducing the incidence and severity of the same. The data was collected from multiple sources recorded between 2000 and 2020 inclusive, then cleaned and prepared, model trained, model tested and validated. Artificial Neural Network

(ANN) was implemented to establish a prediction model that was made accessible through a web interface.

ANN was used since it best met the needs of this study. The solution required an approach that could model complex relationships between inputs and outputs to find patterns in data. ANN is designed to extract existing patterns from noisy data thus was found suitable for this study. This is a study of a real life situation and could benefit from the ability of the ANN to learn and model non-linear and complex relationships that enables it to generalize and predict on unseen data. In addition, ANN does not impose restrictions on the input variables and has proven appropriate for data with high volatility and non-constant variance due to its ability to learn hidden relationships without imposing any fixed relationship in the data.

1.2 Problem Statement

The world is witnessing frequent wild fires many of which are severe and destructive. The fires are burning more acres of land and there is a steady increase of the mega fires occurring due to the rising global temperatures. For instance, Mount Kenya experienced over 100 forest fires in 2019 and majority of the reported fires were caused through human activities and weather factors (Henry, McCarty & Maingi, 2020). A number of fires were also reported in the Amazon in 2019 as well as Australia. The increasing wildfires are steadily decreasing the forests' ability to support economic, recreational, and subsistence activities (Patel, 2018). Protecting and restoring forests has never been more urgent. Proactive management of forests can help in minimizing wildfires by predicting future fires and focusing on preventive efforts thus mitigating the impacts. However, most of the existing wildfire prediction solutions do not take into account the complete data that captures all the necessary factors and components that influence wildfire ignition. Past studies in wildfire predictions have paid less attention to the effects of human factors but given more consideration to climate related and topographical variables to predict wildfire occurrences (Guo, et al., 2016). The use of impartial data can lead to inaccurate prediction of wildfire occurrence resulting to poor decision-making in the management of forest resources.

Based on the aforementioned shortfalls, this study attempted to capture all the factors and components such as weather (temperature and precipitation), topography (elevation and slope), forest type, firebrands, fuel types and human activities to model a solution that correctly predicted wildfire occurrence and provided necessary information to be used by all the stakeholders to create awareness in the need to prevent forest fires.

1.3 Aim

The aim of this project is to develop a supervised machine-learning model to predict wild fire occurrences at different spatial scales through different periods of time using existing data.

1.4 Specific Objectives

i. To investigate the challenges faced in the prediction of wild fires
ii. To investigate existing techniques and models for wild fire prediction
iii. To develop a supervised machine learning model to predict wild fire occurrence
iv. To test and validate the developed system

1.5 Research Questions

i. What challenges do stakeholders encounter in predicting wild fires?
ii. What are the existing techniques and models for wild fire prediction?
iii. How can a supervised machine-learning model to predict wild fire occurrence be developed?
iv. How can the developed system be tested and validated?

1.6 Justification

Climate change is increasing the dangers posed by wildfire across the globe, at the same time rapid human activities around forest areas is providing more sources of ignition thus a big threat to the forest ecosystem. Globally, the extreme weather is becoming a norm and this is not good news to our landscape. There is need to focus on reducing wildfires rather than waiting until fires occur to fight them (Flannigan & Jones, 2020). This research will greatly contribute to the protection of the world forest through accurate prediction leading to reduction of wild fires across the world. The proposed technique will provide a platform for the stakeholders to manage the forests in a sustainable way and able to be effectual in preventing forest fires.

1.7 Scope and Limitation

The primary focus of the research is to propose supervised machine learning model that will give stakeholders an advantage for implementing efficient preventive strategies and measures through utilization of machine learning and predictive analytics.

Chapter 2: Literature Review

2.1 Introduction

This chapter consists of review of the relevant literature that is necessary in assisting to understand the subject of the study. Traditionally, the focus has been placed on the wild fire detection techniques. Unfortunately, these techniques have suffered major blows because of the remote location of the forests resulting to untimely detection and eventual destruction to the forest ecosystem (Aseko, 2018). In order to refocus efforts towards prevention rather than detection, a variety of methods have been used to predict wildfire phenomena and predictive models have exploited several sources of data describing fire phenomena (Taylor, Woolford, Dean & Martell, 2013). Prediction of wild fire occurrence plays a major role in resource allocation, mitigation and recovery efforts. The chapter will focus on providing an understanding of the information needed for fire management decision-making and the challenges involved in predicting fire occurrence. The challenges associated with the prediction of wild fires are investigated and a better way of addressing the issue has been proposed through the use of machine learning.

2.2 Challenges Stakeholders Encounter in Predicting Wild Fires

2.2.1 Weather Effects

Weather performs a central role in the start, development and death of a wildfire. The timing of fires depends on their causes. In the isolated wooded regions, the timing of lightning-caused fires is more linked to weather settings and the season, with majority of such fires happening in summer (Ganteaume et al., 2012). Extreme temperature, protracted drought phases and strong winds play a substantial part in the life cycle of wild fires. According to Bonsor (2001), there are three weather ingredients that affect wild fires:

i. Temperature – the heat from the sun creates potential fuel by drying the sticks, trees and underbrush. Ignition chances are high in warmer temperatures.
ii. Wind – has big impact on wildfire as it supplies oxygen and cannot be easily predicted

iii. Moisture – slows the fire down more so when in humidity and precipitation form thus reducing its intensity. Rain and other precipitation raise the amount of moisture in fuels eventually suppressing any possibility of wildfire breaking.

It is the complexity in understanding the effects of the weather ingredients that poses the main challenge to stakeholders in the prediction of wild fires. There is need to include all the weather parameters for efficient and accurate weather prediction in regards to preventing wild fire occurrence.

2.2.2 Fuel Effects

Wild fire requires fuel such as sticks, trees, underbrush, and dry grass among others to burn. The bigger the fuel load, the higher the intensity and vice versa. The various fuel elements do affect the behaviour of fire. According to Bracmort (2013), the following features of forest fuels affect the nature, spread and intensity of wild fire:

i. Moisture content – moist fuel leads to slow combustion and vice versa. The moisture level of fuel is interlinked to the weather patterns over a period of time.

ii. Size – ignition is faster on small fuels such as dry grass, dead leaves and twigs than the larger ones such as large stems, stumps and logs. The effect of size on fire behaviour is represented in Table 2.1.

iii. Load or amount - this is the amount of fuel available in an area. Low volume causes low intensity fire and vice versa

iv. Arrangement – scattered fuel ignites faster and produces intense fire as they dry faster and get abundant oxygen to facilitate combustion process. The opposite is true for compacted fuels.

v. Shape – it determines ignition ability as well as fire intensity. For example, fuels that are flat dry out quicker and provides bigger surface to absorb heat thus taking less time to dry out.

Table 2. 1: Fuel Size and how it affects fire behaviour (Adapted from (Bennett, 2017)).

Fuel category	Diameter (in.)	Description	Impact on fire behavior
1-hour	0.00–0.25	Needles, twigs, moss, lichens, small shrubs and grasses	Easily ignited. Supports initial fire spread and the heating and combustion of larger fuels. Under dry conditions, these fuels are flashy and surface fires spread quickly.
10-hour	0.25–1.00	Small branches, shrubs	Supports fire spread and the heating and combustion of larger fuels. Under very dry conditions fires spread quickly.
100-hour	1.00–3.00	Medium-size branches	Supports fire spread and the heating and combustion of larger fuels.
1,000-hour	3.00–8.00	Large branches, small logs	Supports fire spread. Increases fire duration and influences fire severity, depending on loading.
10,000-hour	> 8.00	Large downed logs that are solid or moderately decayed	Ignites after flaming front has passed. Large fuel pieces do not support fire spread, but can increase fire duration and severity near the log. If fuel loading is high and distributed across the site (such as from beetle-killed trees), high fire severity can be more widespread and increase resistance to control and the duration of burning.
Snags	Variable	Bole only or bole with large branches, depending on snag condition	When snags combust, they can torch, lofting embers and firebrands ahead of the main fire, starting additional spot fires. Snags may increase resistance to control.

2.2.3 Effects of Topography

Topography captures the surface configuration including its slope and aspect. According to Mickleburgh (2006), the topographic outcomes of slope affect the spread of fire to neighbouring spaces; fires are more likely to spread to uphill areas and less likely to spread downhill. It also affects fire intensity and if by any chance there is fire outbreak from the very top of a slope, the burning process will be slower compared to the one started at the bottom of a slope as captured in the Figure 2.1.

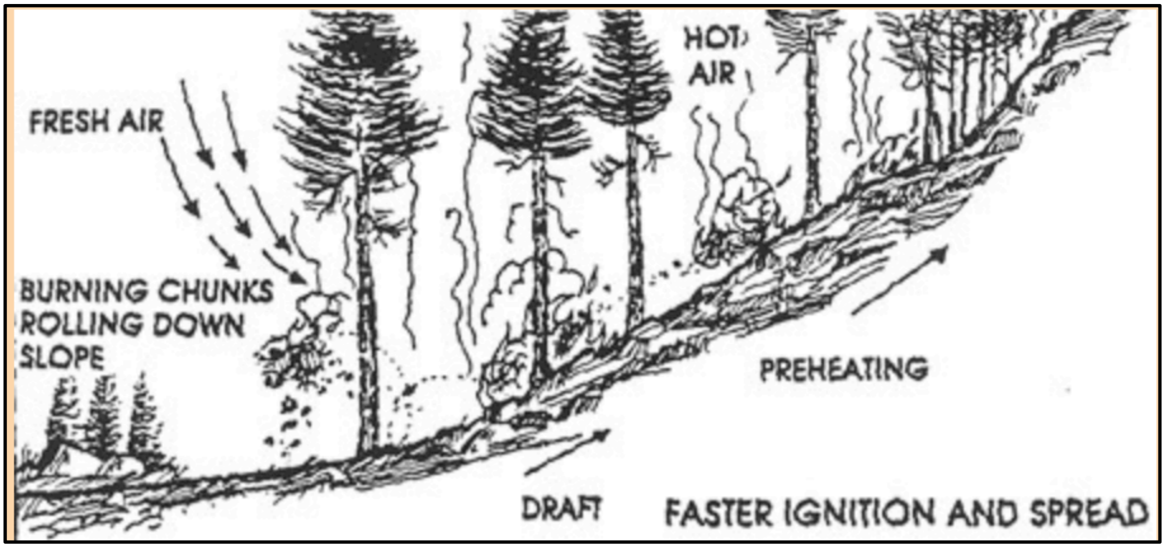

Figure 2. 1: Slopes Intensifies Radiant Heat Transfer and Updrafts (Topography's effect on Fire Behavior, n.d.).

The above, weather, fuels and topography form the fire behaviour triangle. It is challenging to evaluate the net effect of these factors in predicting wild fires. There is need to model a system that can effectively evaluate how weather, fuels, and topography impede fire spread in forested areas and how to use the same system for stakeholders to proactively respond to emerging and on going wildfire threats.

2.2.4 Effects of Human Activities

Most of the wild fires across the world are related to human activities. Human factors ranging from structures built in forest areas to ignition sources such as cigarettes on the highway to faulty electric poles significantly contributing to the wildfires and that about 90 per cent of wildfires recorded in California are due to human activities (Mann et al., 2016). Camp and Dorflinger (2018) identified the following types of ignitions that they attributed to about 84 to 95 per cent of fires throughout the United States as depicted in the Figure 2.2.

i. Direct Human Causes – Arson, debris burns, campfires, and cigarettes
ii. Indirect Human Causes – Power lines and equipment sparks.

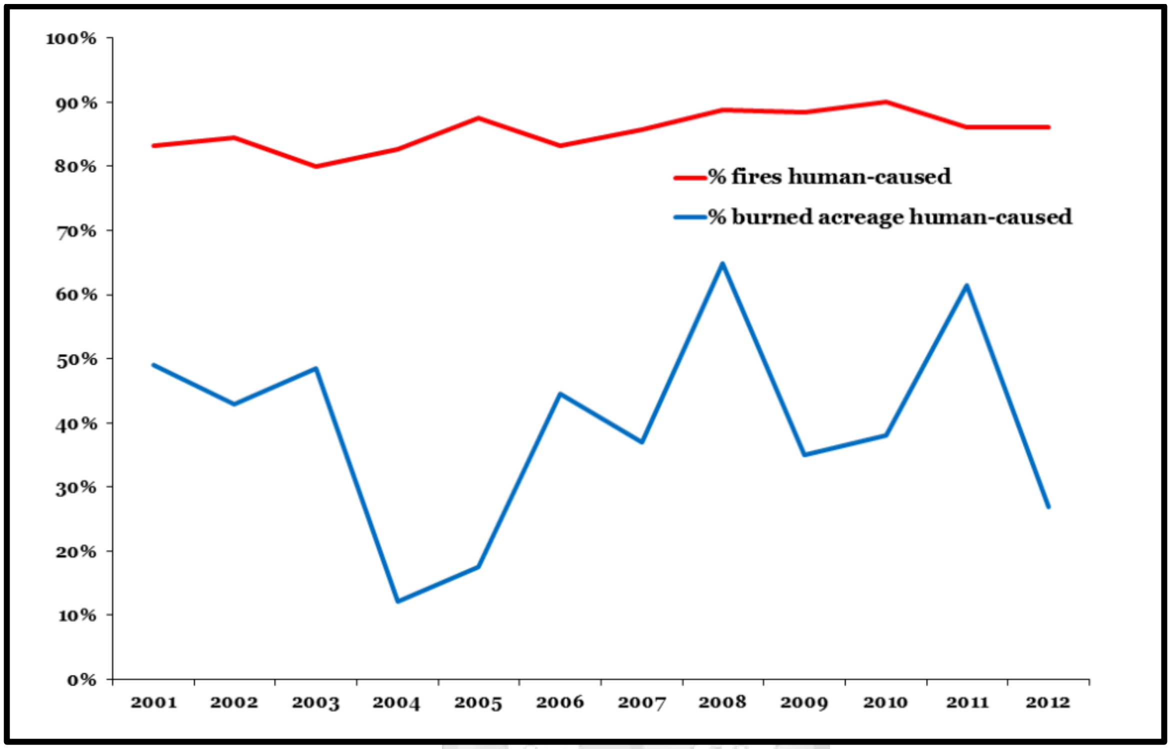

Figure 2. 2: Cause of Wildfires in United States between 2001-2012 (Tobin, 2013).

There is need to understand the relationships between human activities and wildfires with an aim to accurately account for them. This has been a major challenge in predicting wild fire occurrence, as it is crucial to consider both climate and human variables when modelling wild fire prediction system.

2.3 Existing Techniques And Models for Predicting Wildfires

The models and techniques for wildfire predictions differ from one region to another as governments, international agencies and other stakeholders entrusted with the management of forests have failed to agree on how fires should be managed (Moore, Hardesty, Kelleher, Maginnis, & Myers, 2003). These models have the ability to predict wildfire events helping stakeholders to make key decisions aimed at mitigating and reducing their impact. Among the models of wildfire prediction is the data-driven approach that uses weather data as well as the use of indexes to measure fire potential in wild lands. Equally there is the probabilistic method that predicts fire probability by introducing an Ignition Component (IC) to the weather data.

2.3.1 Context-Based Fire Risk (CBFR) Model

This is a dynamic data-driven wildfire danger prediction model that uses weather data to address the above-mentioned challenges using context-based anomaly detection techniques (Moore, Hardesty, Kelleher, Maginnis, & Myers, 2003). Moore et al., (2003) achieved this through the use of unsupervised algorithm and weather data based on meteorological observation measuring humidity (H) in %, air temperature in ^{o}C, average wind velocity in the open at the height of 10 m in km hr^{-1} and drought factor (DF) which uses precipitation observations Pre. The advantage of this model is that it supports several historical models for different temporal variations (for example day versus night) and uses ensemble-learning techniques to predict wildfire risk with high accuracy. Additionally, its unsupervised nature makes it flexible and does not rely on expert knowledge hence it can be deployed at any area of interest.

However, Context-Based Fire Risk (CBFR) Model has the following shortcoming; it only considers the weather factor to predict wildfire yet humans are responsible to most of the wildfires around the world. For more accurate predictions, it is critical for a model to proportionately account for weather factors and human activities (Mann et al., 2016). Context-Based Fire Risk (CBFR) Model therefore uses partial data thus its prediction is inaccurate.

2.3.2 McArthur Forest Fire Danger Index (FFDI)

FFDI model is widely used to forecast the influence of weather on fire behavior and is based on the temperature T (^{o}C), wind speed v (km h^{-1}), relative humidity RH (%) and Drought Factor (DF) representing fuel availability (Dowdy, Mills, Finkele & Groot, 2009). McArthur Forest Fire Danger Index also uses climatological components only to create a weather index of fire potential. The produced index is converted into fire categories used to warn the stakeholders as per the Figure 2.3.

Figure 2. 3: Forest Fire Danger Rating System (Adapted from (Doyle, Deacon & Locke, 2017)).

Despite being wildly used, FFDI have some flaws; it does not look at all conditions that impact on fire behaviour such as wind changes as well as human activities that have influenced wildfires for awhile now. It is very sensitive at extreme end of the scale and any minor modifications to the input variables e.g. humidity, wind speed and temperature have a big influence on the fire danger index.

2.3.3 The National Fire Danger Rating System (NFDRS)

NFDRS offers indexes that can be used to measure fire potential in the forests (Cohen & Deeming, 1985). Helfman et al., (1980), noted that both Federal and States use this system, and that data from fire-danger rating stations all over USA get administered through the interactive, time-share computer system called AFFIRMS (Administrative and Forest Fire Information Retrieval and Management System). NFDRS incorporates the effects of prevailing and anticipated states of chosen fire danger factors into one or more qualitative or numeric indices that mirror a region's protection need (Schlobohm & Brain, 2002). NFDRS considers a combination of fuel, weather and topography in predicting fire occurrence possibility. Schlobohm and Brain (2002) identified the following key components of a fire danger rating system:

i. Model representation of the connection between fuels, weather and topography and the their influence on fire matters.
ii. A system to gather data necessary to produce the rating numbers
iii. A processing system to convert inputs to outputs and perform data analyses.
iv. A communication system to share the fire danger rating information between entities.
v. A data storage system to retain data for historic reference.

The main advantage of NFDRS is the ease of centrally storing information and sharing the same. The weakness of this approach is that it does not consider firebrand i.e. the pieces of material dropped by humans (Sun & Zhang, 2018). Gossner and Chief (1999) also noted complexity as one of the major drawback to the NFDRS due to its continuous development to keep up to date. The complex nature of this system makes it difficult to understand thus the need to spend more time on training various stakeholders to understand it. Figure 2.4 shows the structure of NFDRS.

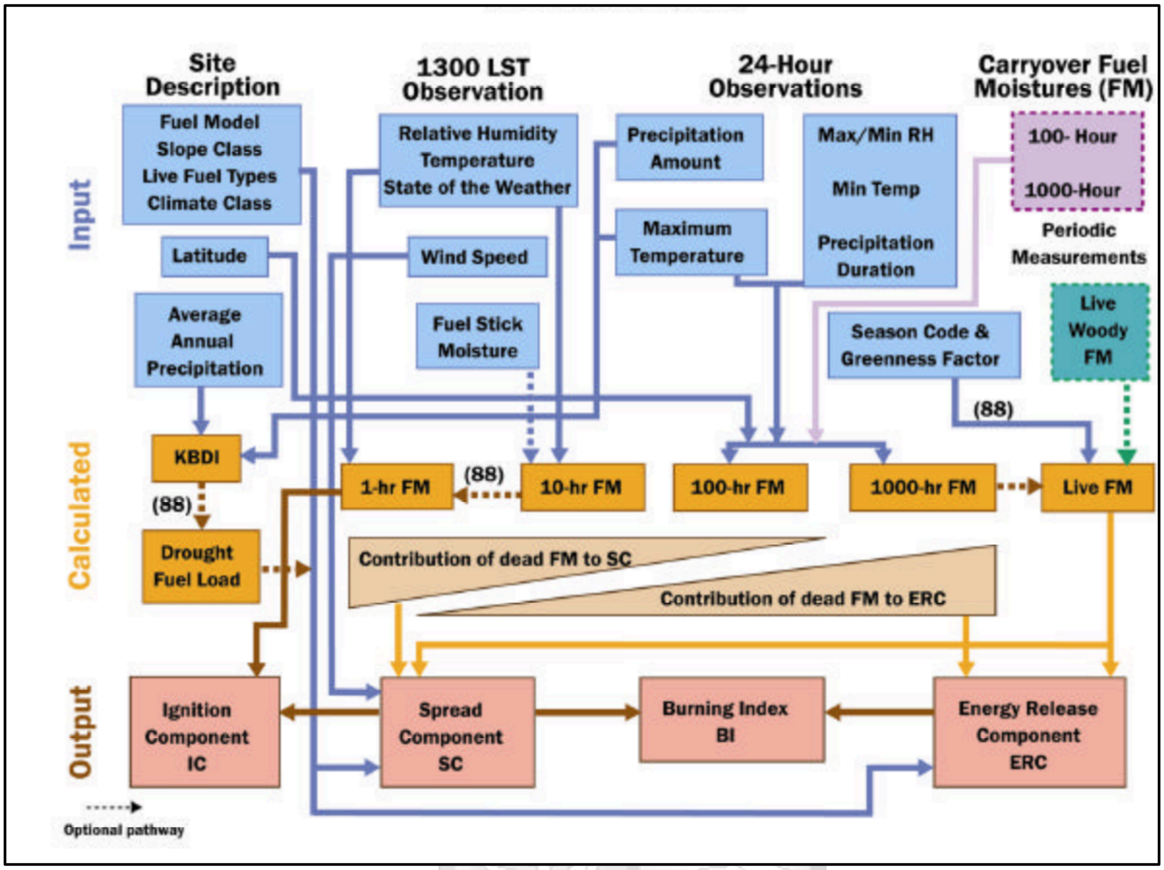

Figure 2. 4: National Fire Danger Rating System Structure (Adapted from (Schlobohm & Brain, 2002)).

2.3.4 A Probabilistic Method Predicting Forest Fire Occurrence

This is an enhancement to the NFDRS through the inclusion of an Ignition Component (IC) (Sun & Zhang, 2018). The system is currently being used in the Northern part of the Daxinganling Region, China. Weather, firebrand and fuel affect the probability of fire occurring. Sun and Zhang (2018), expressed fire probability as a product of firebrand probability, fuel ignition probability and the probability of ignition spreading to a reported fire. The advantage of this model is that it provides a fire danger alert that minimizes the inconsistency between fire danger rating prediction and actual fire occurrence probability. On the other hand, it shares some of the limitations observed on NFDRS such as complexity that makes it difficult to understand.

2.4 Machine Learning Models

Robert (2014) defines machine learning as a set of methods that can automatically detect patterns in data and the use the uncovered patterns to make decisions. There are various machine learning models for prediction built using different algorithms. Below are some of the algorithms:

2.4.1 Artificial Neural Networks (ANN)

Artificial Neural Networks have been used in wild fire predictions and prevention. Sayad, Mousannif and Moatassime (2019) used this algorithm to process data collected from various sources, extracted insights from them to predict the occurrence of wildfires. Their investigational outcomes provided high predication accuracy. They validated the model using classification metrics, regularization, cross-validation and model comparison. Siva (2019) observed that ANN has self-learning abilities and this gives it the ability to produce (output) better outcomes, as more data is input as illustrated in Figure 2.5. Neural networks are fine-tuned based on assessment of the output and the target, until such a time that the network output matches the target. ANN has three interconnected layers with the first one being the input, while the second and third being the output.

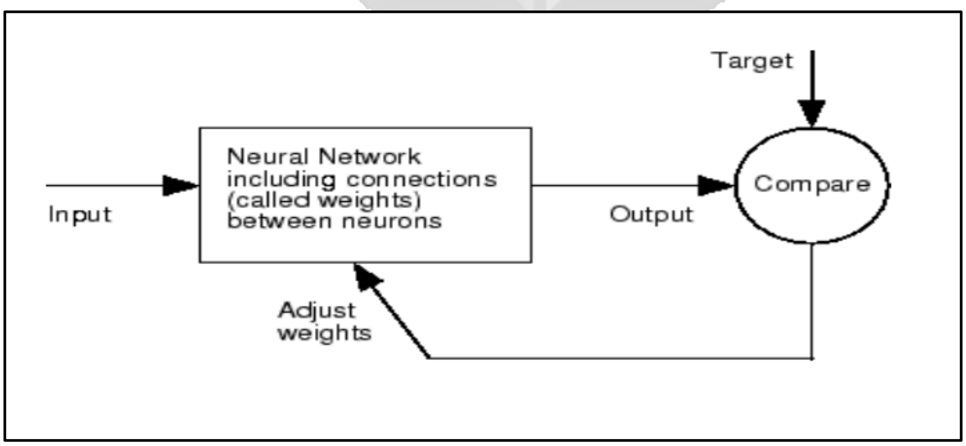

Figure 2. 5: Structure of ANN (Adapted from (Dey, Bhoumik & Dey, 2016))

2.4.2 Support Vector Machine (SVM)

SVM belongs to the supervised algorithm learning class. It can solve both linear and non-linear problems by creating a line or a hyper plane, which separates data into classes (Pupale, 2018). SVM hyper planes are determined by subset of training instances called support vectors. Figure 2.6 by Manning et al., (2009) illustrates the same.

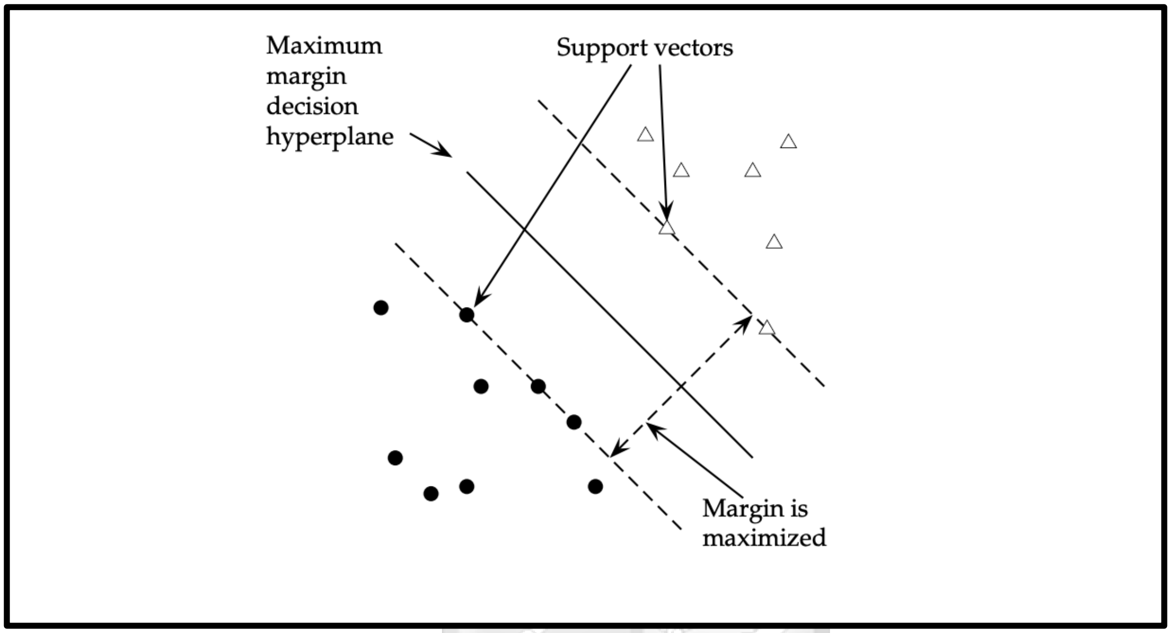

Figure 2. 6: Support Vector Machine (Adapted from (Manning et al., 2009))

Sukumar et al., (2009) deployed this algorithm to predict forest fire using temperature and humidity values prevailing inside the forest. The values were given to the learned SVM classifier to predict the class labels. A class prediction of 'YES' indicated high chance of fire occurrence, thus calling for preventive measures to be put in place. In their paper "Wildfire Predictions: Determining Reliable Models using Fused Dataset", Naganathan et al., (2016), declared SVM as the best optimum models for binary and multicast classifications on the selected fused dataset after evaluating K-Nearest Neighbors and decision tree models as well.

2.4.3 K-Nearest Neighbors (KNN)

Harrison (2019) describes KNN as a simple, easy-to-implement supervised machine-learning algorithm for solving both classification and regression problems. Naganathan et al. (2016), while evaluating the reliability of three predictive methods i.e.

Support Vector Machine, K-Nearest Neighbors and decision tree models using fused dataset in wildfire prediction, noted that K-Nearest Neighbor (KNN) had demonstrated to be timely, cost-efficient, and accurate when applied to Northern European countries and the US. Rishickesh, Kumar, Shahina and Khan (2019) refer to KNN as an instance-based machine-learning algorithm specialised in local approximation of a function. abarty and Bhattacharya (2015) noted that K-NN is a very good classifier that can classify two classes of events 'storm days' and no 'storm days' and the same can be adopted in wildfire prediction.

2.5 Literature Review Summary

The chapter has presented literature review on the research topic based on the objectives of investigating the challenges faced in the prediction of wild fires and investigating existing techniques and models for wild fire prevention. It also evaluated some of the relevant machine learning models and the development of a machine-learning tool for wild fire prediction. The research work builds on the existing studies done on wild fire prediction and prevention.

2.6 Conceptual Framework

Figure 2.7 illustrates how the model works. Raw data from various sources is collected then stored in a file. The raw data is then transformed into an understandable format. The prepared data is then used to train machine-learning algorithm to give predictions on the possibilities of wildfires occurring.

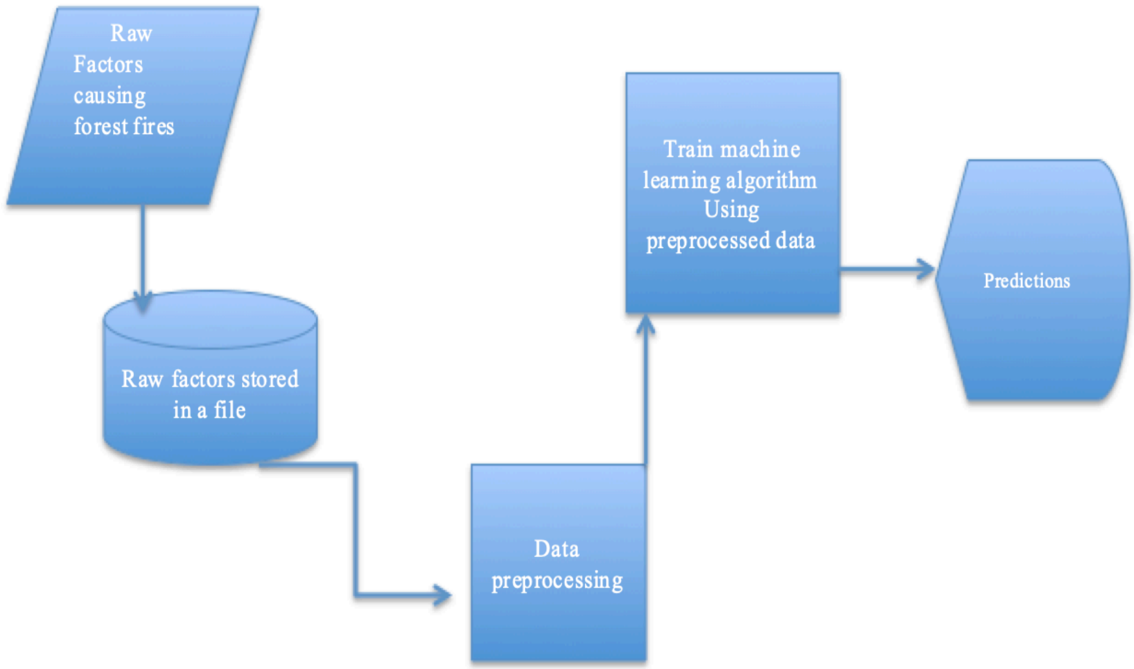

Figure 2. 7: Conceptual Framework

Chapter 3: Research Methodology

3.1 Introduction

This chapter presents the research methodology used to determine the data features influencing the prediction of wild fires, the machine learning model selection approach and the implementation of wild fire prediction model. Kothari (2004) define research methodology as a systematic way to solve research problem in a way that outlines the various steps adapted by the researcher in studying a research problem along with the logic behind them. This chapter aims to discuss the following subtopics: research design, model development (data collection, data pre-processing, model training and model testing and validation), system development methodology, and ethical considerations.

3.2 Research Design

The study implemented an applied research aimed at solving a critical and real-world problem faced by forest management stakeholders by developing a model to be used in wild fire prediction. The study identified the research objectives and built a web-based model to prove the concept.

3.3 Model Development

The model development took the following steps:

i. Data Collection
ii. Data Pre-processing
iii. Model Training
iv. Model Testing and Validation

3.3.1 Data Collection

The main source of data was from climatological data that includes meteorological (weather) data as well as fire databases such as National Fire Danger Rating System (NFDRS), University of California machine learning repository, and Canada National Fire Database (CNFDB). The data was reflective of conditions experienced or anticipated to occur within identified forests (Naganathan et al., 2016). Some data were also collected from scientifically verified Internet sources, authentic figures in Forest management, and reviews and documentation by organizations such as Food and Agriculture Organization (FAO), World Wide Fund for Nature (WWF), and/or Forest Management authorities. Moreover, review of related past studies on the subject were conducted for data collection and identification of the existing gaps.

The data collected was mainly from Canada, USA and Portugal. A total number of 20000 data volumes were collected and out of which 517 records were used in a dataset. Each record had ten parameters that included moisture content of litter, moisture content of loosely compacted organic layers, moisture content of deep compacted organic material, temperature, relative humidity, wind speed, rain, land terrain, fire causes and occurrences.

3.3.2 Data Pre-processing

The collected data was extracted from their various sources and stored in a worksheet as Excel documents. The needed data were identified while corrupt and/or inaccurate records were detected and corrected (removed). Complete data was saved for training, testing and validation of the model. The aim was to ensure that the model produced was valid and reliable.

3.3.3 Model Training

The collected data was represented in a document-term matrix with tf-idf feature weighting to enable the application of ANN machine learning algorithm. The collected data were spilt into 80% training sets and 20% testing sets with an aim to train and validate the model.

3.3.4 Model Testing and Validation

A number of tests were conducted to establish if the developed model satisfied the proposed requirements. Sample dataset was set aside for validation with an aim to give an unbiased evaluation of the model.

3.4 System Development Methodology

The model was developed following the Rapid Application Development (RAD) system development methodology, as it needed to be built incrementally until it was finished within a short time (Kapaya, 2017).

3.4.1 System Analysis

This phase involved much on requirement gathering such as requirement specifications, hardware analysis, and software analysis.

3.4.2 System Design

The design approach deployed was rapid application development (RAD). The output from this phase was expected to be the system architecture diagram, the system sequence diagrams, and the database schema diagram.

3.4.3 System Implementation

This is the phase where the system for predicting wild fire occurrence was built; the implementation took three main steps as follows, construction, installation and maintenance.

3.5 Ethical Considerations

The researcher ensured that any information retrieved from various sources was mainly used for the purpose of this research. This study involved referral to works from other researchers and authors, which were documented and credit given to such works to avoid plagiarism.

Chapter 4: System Design and Architecture

4.1 Introduction

This chapter details the design and architecture of the web-based wild fire occurrence predication system that can be used by various stakeholders involved in the management of forests and wild lands based on the conceptual model presented in Figure 2.7. The chapter helps clarify the functionality of the developed system and the iteration between different components.

4.2 Requirement Analysis

This phase involves the review of users' expectations so as to ensure that the model takes into account all the stakeholders' needs as per the set objectives of this research study. On the same note and in accordance to the set objectives, the discussion below is mainly on the functional and non-functional requirements of developed model.

4.2.1 Functional Requirements

Functional requirements describe what the model has to do by pinpointing the tasks, actions and/or activities that must be accomplished. The functional requirements of the model include:

i. The system should allow for users to self-register on first visit.
ii. The system should allow for uploading of a csv file into a folder.
iii. The system should enable uploading of raw dataset for training and testing.
iv. The system should have search capabilities for prediction.
v. The system should display prediction data using text.
vi. The system should provide recommendation to the user based on the fire predictions

4.2.2 Non-Functional Requirements

These are the general requirement attributes and qualities that the system must have for optimal performance. The non-functional requirements include:

i. User friendly - The system will have a user-friendly user interface making it easy to learn, operate, prepare inputs and interpret outputs as users interact with it.

ii. Access Security – The system will be safeguarded against deliberate and intrusive faults from internal and external sources
iii. Availability – The system will be dependable and will be able to functional around the clock.
iv. Accessibility – People with the widest range of capabilities will be able to use the system to achieve the specified study objectives.
v. Confidentiality – The system will protect sensitive data and allow only authorized access to the data.
vi. Efficiency – The system will be able to handle capacity and throughput within the specified maximum time.
vii. Integrity – The system data will be maintained accurately, authentically and without corruption.
viii. Maintainability – The system will allow for ease of finding faults and errors in order to fix them to meet the user requirements.

4.3 System Architecture

The system architecture for the wild fire occurrence prediction system is shown in Figure 4.1. It illustrates the general interaction of various components to achieve system functionality. Raw data is pre-processed to create training and testing datasets. Thereafter, the Artificial Neural Network is trained and validated then converted to a format that can be embedded into a web application. The web interface displays predictions as per users' requests, at the same time it allows for users to input their queries.

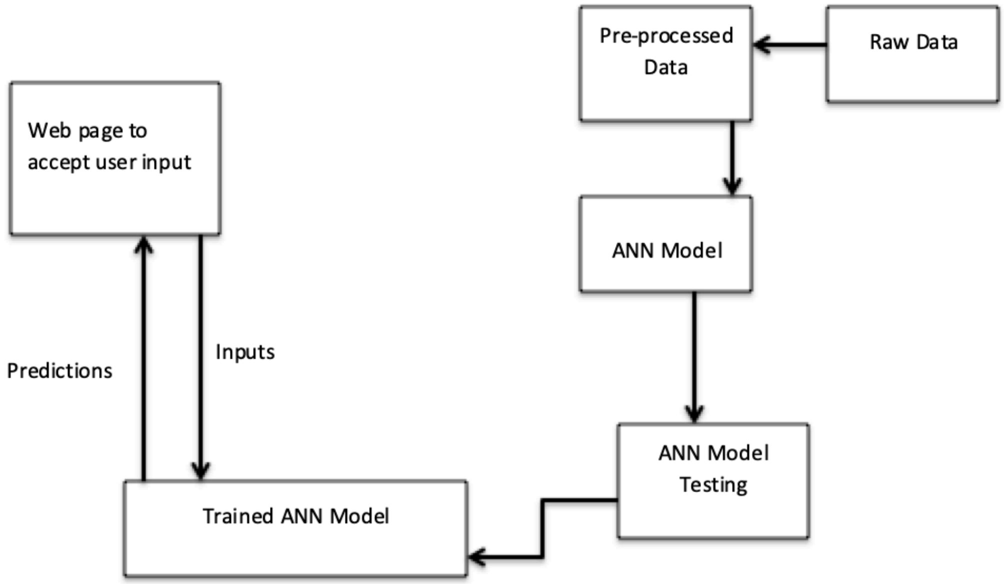

Figure 4. 1: System Architecture

4.4 Use-Case Diagram

Use case diagram demonstrates actions and event steps between actors and the system. It represents the users interacting with the system. Figure 4.2 depicts the functionality that the proposed model should have. The main systems actors in here are the users and system administrator. The System Administrator collects, cleans the data then trains the artificial neural network algorithm and finally tests the artificial neural network model to discover data and knowledge presented to the user through web interface.

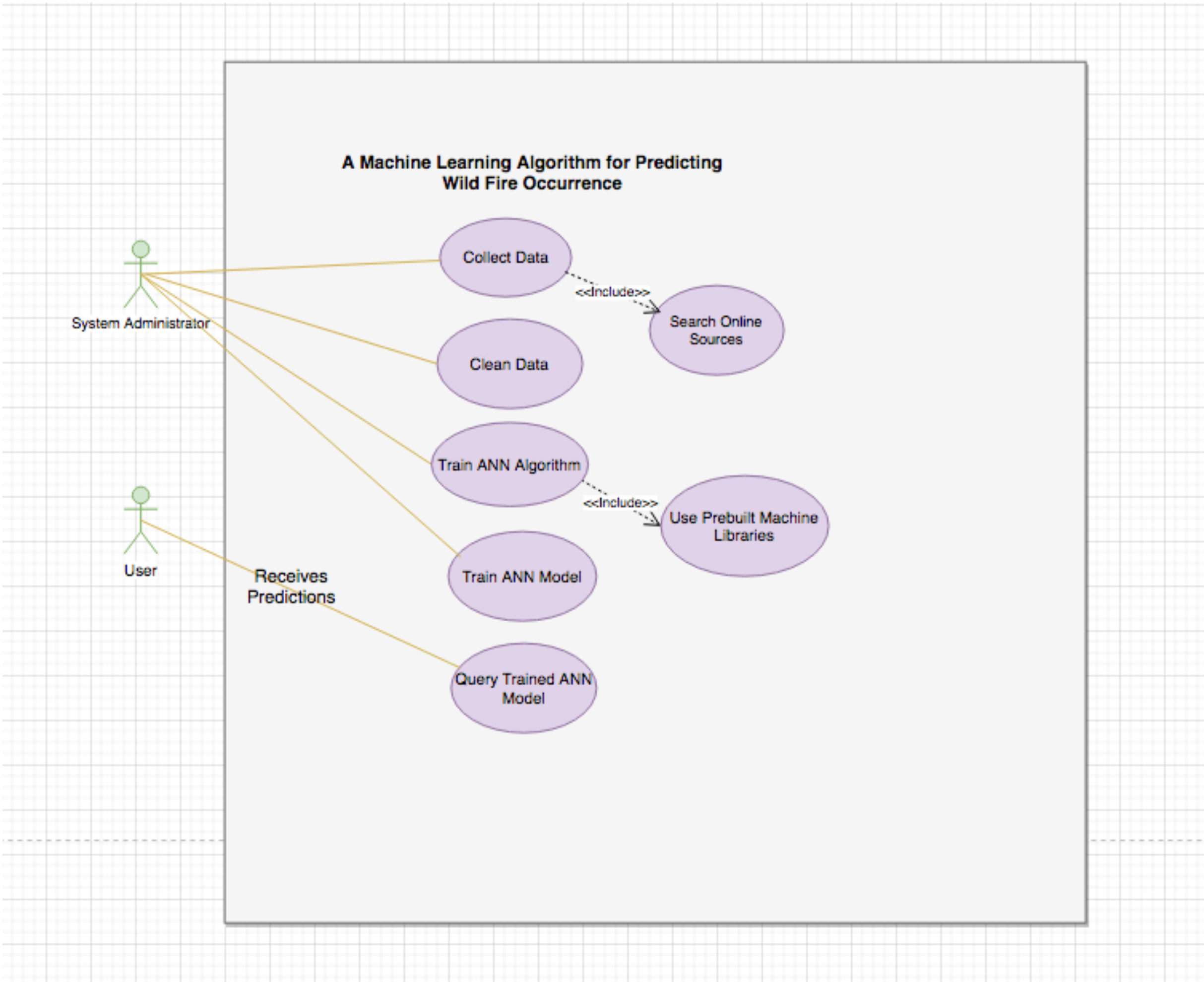

Figure 4. 2: Use Case Diagram

Table 4. 1: Data Collection Description

Use Case: Collect data	
Primary Actors: System Administrator	
Brief Description: This use case describes how the System Administrator will collect the data	
Pre-condition: Data is relevant to the causes of wildfire	
Post-condition: System Administrator identifies the sources and retrieves the data.	
Main Success Scenarios	
Actor Responsibility	**System Responsibility**
1. The system administrator identifies the relevant sources.	
2. System administrator retrieves and collects the identified data	
	3.The system stores and saves the collected.

Table 4. 2: Clean Data Description

Use Case: Clean Data	
Primary Actors: The System Administrator	
Brief Description: This use case describes how the System Administrator will clean the raw dataset	
Pre-condition: Adequate data available	
Post-condition: Pre-processed dataset	
Main Scenarios	
Actor	**Systems**
1. The system Administrator imports the libraries	
2. The System Administrator imports the data-set	
3. The system Administrator checks out the missing values	
4. The system administrator splits the data-set into Training and Test Set	
	5. The System performs feature extraction from
	6. The System saves the extracted features

Table 4. 3: Model Training Description

Use case: Model Training	
Primary Actors: The System Administrator	
Brief Description: This use case describes how the System Administrator will train the Artificial Neural Network Algorithm	
Pre-condition: Pre-processed dataset, training system available	
Post-condition: Trained model that can predict likelihood of wildfire occurrence	
Main Success Scenarios	
Actor	**Systems**
1. Administrator selects pre-processed data	
2. Administrator selects training and testing set percentage	
3. Administrator selects output format	
4. Administrator executes training command	
	5. System splits dataset into training sets as executed by the administrator
	6. System outputs the trained model

4.5 System Sequence Diagram

The diagram illustrates the interaction between the user and the trained system. The user inputs prediction parameter they are interested in through the user interface, then query the model. The prediction out is displayed to user through the user interface. The proposed system's sequence diagram is illustrated in Figure 4.3.

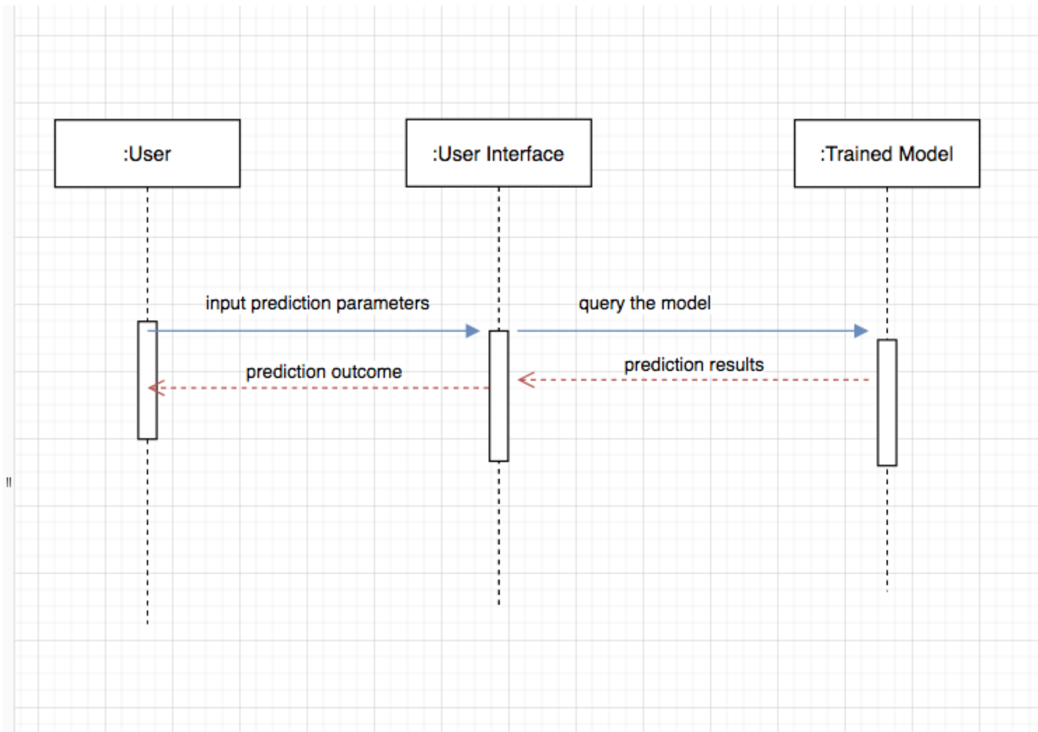

Figure 4. 3: Sequence Diagram

4.6 Context Diagram

4.6.1 Context Diagram Level 0

The Context Diagram defines the system's boundaries identifying the information flow between the system and the external entities. The main entities interacting in the proposed system are user and the trained ANN model. The context diagram shown in Figure 4.4 illustrates the model interaction with the external entities.

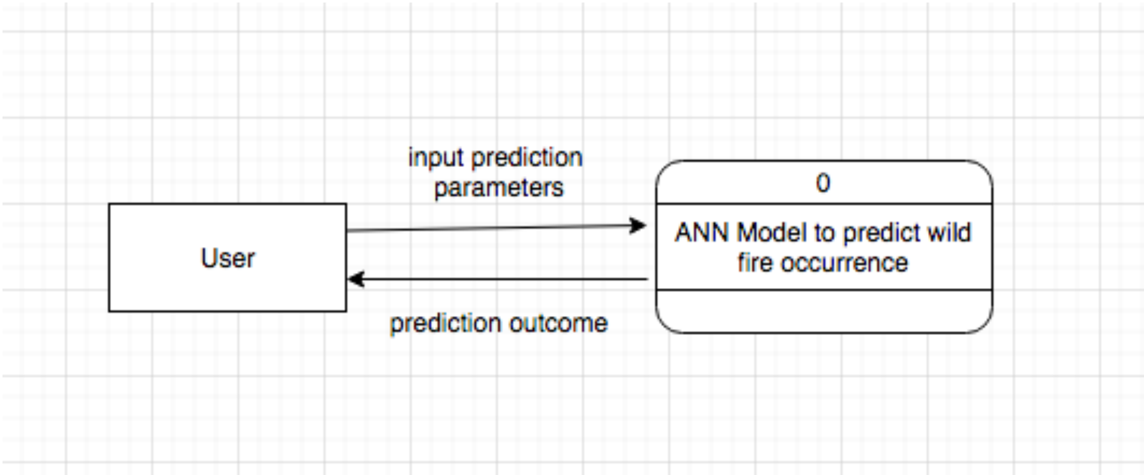

Figure 4. 4: Context Diagram

4.6.2 Context Diagram Level 1

The level 1 context diagram depicted in Figure 4.5 clearly represents the processes, entities, and data stores in the proposed system enabling a clear understanding of the data flow process.

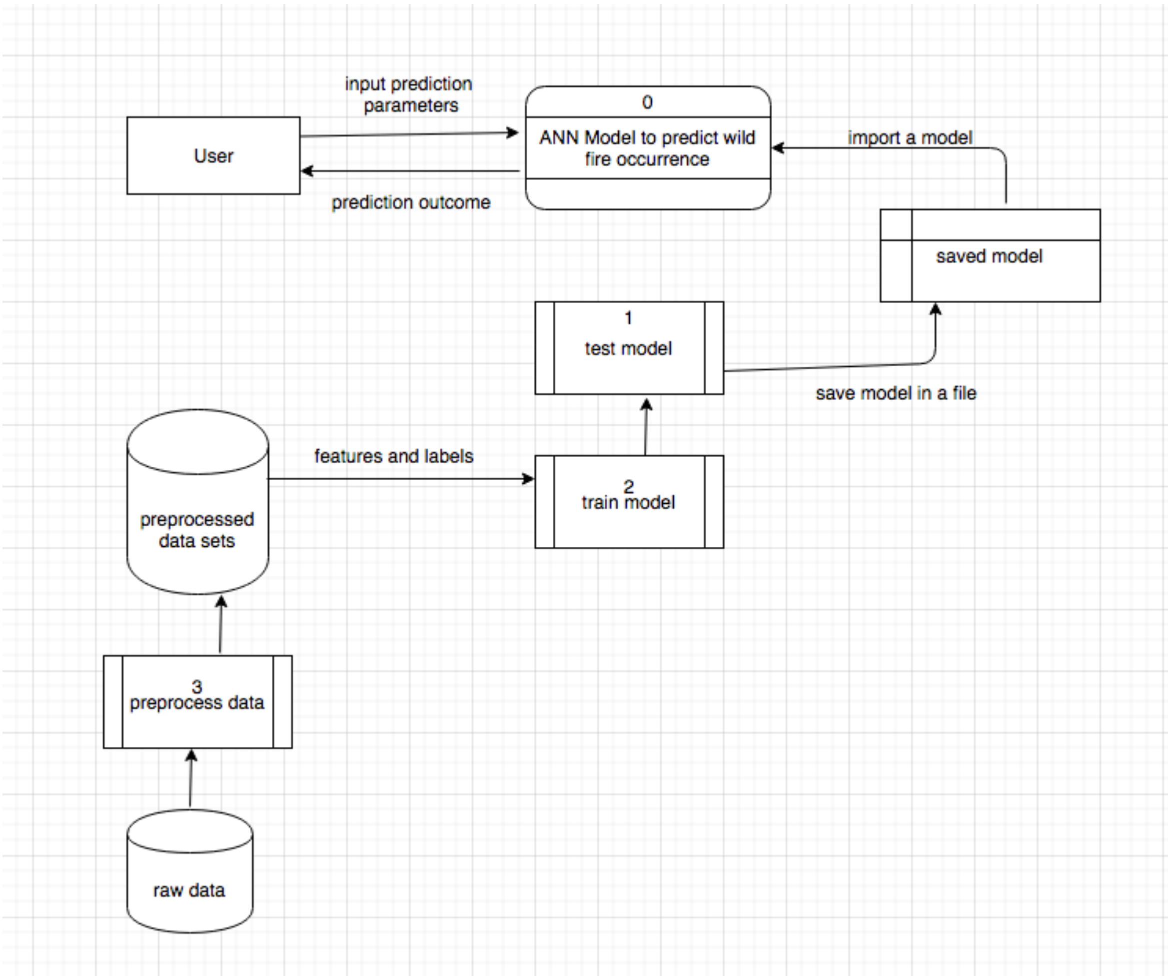

Figure 4. 5: Context Diagram Level 1

Chapter 5: System Implementation and Testing

5.1 Introduction

This chapter details the actual system implementation, testing and validation. The chapter explains the process involved in building the machine-learning algorithm for predicting wildfire occurrence. The chapter also discusses the pre-processing phase as well as the training of the model. The model was then tested against the test dataset obtaining an accuracy of 82.69%. Finally, the chapter describes the use of the model in predicting wildfire occurrences.

5.2 Building the Features

Raw data was collected from various scientifically verified Internet sources as illustrated in Figure 5.1 and 5.2.

Export to CSV

Agency - Any - | Stage of Control - Any - | Response Type - Any - | Apply

Agency Fire ID	Agency Code	Agency Data Timezone	Situation Report Date	Status Date	Stage of Control Status	National Fire Cause	Response Type	Fire Size	Latitude	Longitude
2020-R40115	BC	PDT	2020-04-20 17:59	2020-04-20 23:24	BH	H	FUL	4.0	55.62662	-128.11983
2020-K50120	BC	PDT	2020-04-20 21:14	2020-04-20 23:23	OC	H	FUL	0.1	49.90583	-119.59800
2020-G80109	BC	PDT	2020-04-19 21:45	2020-04-20 23:16	OUT	H	FUL	1.0	56.58280	-120.49020
50646	NB	ADT	2020-04-19 18:26	2020-04-20 23:00	OC	U	FUL	0.8	45.62370	-65.54020
50703	NB	ADT	2020-04-20 13:34	2020-04-20 23:00	OC	U	FUL	0.1	45.65363	-65.67398
50742	NB	ADT	2020-04-21 06:51	2020-04-20 23:00	OC	U	FUL	1.0	45.54195	-66.57222
50707	NB	ADT	2020-04-20 15:18	2020-04-20 23:00	OC	U	FUL	0.1	45.60633	-64.94667
2020-G80116	BC	PDT	2020-04-20 19:52	2020-04-20 22:24	OC	H	FUL	200.0	56.23612	-121.45378
2020-G50114	BC	PDT	2020-04-20 16:05	2020-04-20 20:53	OUT	H	FUL	1.0	54.55235	-124.42295
2020-N10119	BC	PDT	2020-04-20 20:51	2020-04-20 20:51	OC	H	FUL	0.1	49.51773	-115.73290

Figure 5. 1: Collecting Data from scientifically verified Internet Sources

2018 National Burned Area Composite (shapefile, version 20190919)

2017 National Burned Area Composite (shapefile, version 20190919)

2016 National Burned Area Composite (shapefile, version 20190919)

2015 National Burned Area Composite (shapefile, version 20190919)

2014 National Burned Area Composite (shapefile, version 20190919)

2013 National Burned Area Composite (shapefile, version 20190919)

2012 National Burned Area Composite (shapefile, version 20190919)

2011 National Burned Area Composite (shapefile, version 20190919)

2010 National Burned Area Composite (shapefile, version 20190919)

2009 National Burned Area Composite (shapefile, version 20190919)

2008 National Burned Area Composite (shapefile, version 20190919)

2007 National Burned Area Composite (shapefile, version 20191129)

Figure 5. 2: Collection of yearly-accumulated data from CNFDB

5.3 Dataset Pre-processing

The dataset was retrieved from multiple sources including National Fire Danger Rating System (NFDRS), Canada National Fire Database (CNFDB) and scientifically verified Internet sources. The collected training data was in an unstructured format unsuitable for use in machine learning techniques, thus requiring it to be pre-processed before using it for the training of the model. Figure 5.3 depicts the sample of the collected dataset.

K	L	M	N	O	P	Q	R	
AFEDATE,D	CAPDAT	POLY_HA,N,19,11	ADJ_HA,N,19,11	ADJ_FI	AGENC	BT_GID	VERSIO	COMMENTS,C,254
06/19/16	06/18/16	0.8363734233	0.8363734233	0	QC	169	2016_r9	20161080238[section:intensive,org_source:SOPFEU,n
08/08/16	09/16/16	64.785200043	64.785200043	0	NT	22	2016_r9	Null
	05/10/16	24.1627632676	24.1627632676	0	AB	170	2016_r9	PWF-033-2016[alias:,pt_repdate:2016-05-10,fire_class
	06/26/16	28.9171476597	28.9171476597	0	AB	229	2016_r9	MWF-032-2016[alias:,pt_repdate:2016-06-26,fire_clas
09/08/16	09/25/16	5186.95079045	5186.95079045	0	NT	15	2016_r9	Null
	10/03/16	3.5841688908	3.5841688908	0	AB	230	2016_r9	PWF-115-2016[alias:,pt_repdate:2016-10-03,fire_class
	08/29/16	22.7759182369	22.7759182369	0	AB	232	2016_r9	HWF-245-2016[alias:,pt_repdate:2016-08-29,fire_class
		2.4017385445	2.4017385445	0	BC	175	2016_r9	C10248[track_date:2016-01-01,fire_date:2016-01-01,f
09/30/16	09/15/16	2386.57816069	2386.57816069	0	YT	17	2016_r9	Null
		13.3383390053	13.3383390053	0	NL	172	2016_r9	607]
	07/15/16	13.4738485264	13.4738485264	0	AB	174	2016_r9	HWF-156-2016[alias:,pt_repdate:2016-07-15,fire_class
07/23/16	09/20/16	11.7768002196	11.7768002196	0	NT	10	2016_r9	Null
	09/27/16	1170.8262437	1170.8262437	0	QC	11	2016_r9	Null
	07/07/16	1.5070647551	1.5070647551	0	AB	304	2016_r9	WWF-025-2016[alias:losegun Lake Fire,pt_repdate:20
		4.6736562875	4.6736562875	0	BC	505	2016_r9	C20026[track_date:2016-01-01,fire_date:2016-01-01,f
		1.856847889	1.856847889	0	BC	611	2016_r9	R10023[track_date:2016-01-01,fire_date:2016-01-01,f
09/12/16	09/25/16	2306.84045458	2306.84045458	0	NT	20	2016_r9	Null
	05/20/16	131.748945634	131.748945634	0	AB	612	2016_r9	LWF-076-2016[alias:,pt_repdate:2016-06-27,fire_class
	06/13/16	1655.47108339	1655.47108339	0	AB	861	2016_r9	SWF-030-2016[alias:,pt_repdate:2016-06-13,fire_class
08/02/16	11/13/16	121.825792866	121.825792866	0	SK	137	2016_r9	Null
09/08/16	09/05/16	1596.35257492	1596.35257492	0	NT	138	2016_r9	Null
		317.295407359	353.511261513	1	PC-WB	144	2016_r9_20190919	
		2.265645978	2.265645978	0	BC	159	2016_r9	V70156[track_date:2016-01-01,fire_date:2016-01-01,f
		1.0817070363	1.0817070363	0	BC	866	2016_r9	G40047[track_date:2016-01-01,fire_date:2016-01-01,f
	04/19/16	251.120911432	251.120911432	0	AB	867	2016_r9	PMD-001-2016[alias:,pt_repdate:2016-06-03,fire_class
07/26/16		2.9713394798	2.9713394798	0	MB	404	2016_r9	NE088[fornum:117088,priozone:R,ignidate:2016-07-1
	08/18/16	3786.28145729	3786.28145729	0	BC	59	2016_r9	Null
09/12/16	09/22/16	2289.65320264	2289.65320264	0	NT	58	2016_r9	Null
09/08/16	08/26/16	7106.3840395	7106.3840395	0	NT	38	2016_r9	Null

Figure 5. 3: Sample Dataset

The dataset contained the coordinates of the locations and the occurrence dates of all the fires. It also contained specific information on temperature, relative humidity, moisture content of loosely compacted organic layers, moisture content of deep, compact organic layers, Month of the year, wind speed, and amount of rain. The collected data were processed to only include the parts that were relevant to study. Further processing was done to remove any punctuation. Figure 5.4 illustrates the cleaned dataset.

1	FFMC	DMC	DC	ISI	temp	RH	wind	rain	Causes	Area
2	86.2	26.2	94.3	5.1	8.2	51	6.7	0	0	0
3	90.6	35.4	669.1	6.7	18	33	0.9	0	0	0
4	90.6	43.7	686.9	6.7	14.6	33	1.3	0	0	0
5	91.7	33.3	77.5	9	8.3	97	4	0.2	0	0
6	89.3	51.3	102.2	9.6	11.4	99	1.8	0	0	0
7	92.3	85.3	488	14.7	22.2	29	5.4	0	0	0
8	92.3	88.9	495.6	8.5	24.1	27	3.1	0	0	0
9	91.5	145.4	608.2	10.7	8	86	2.2	0	0	0
10	91	129.5	692.6	7	13.1	63	5.4	0	0	0
11	92.5	88	698.6	7.1	22.8	40	4	0	0	0
12	92.5	88	698.6	7.1	17.8	51	7.2	0	0	0
13	92.8	73.2	713	22.6	19.3	38	4	0	0	0
14	63.5	70.8	665.3	0.8	17	72	6.7	0	0	0
15	90.9	126.5	686.5	7	21.3	42	2.2	0	0	0
16	92.9	133.3	699.6	9.2	26.4	21	4.5	0	0	0
17	93.3	141.2	713.9	13.9	22.9	44	5.4	0	0	0
18	91.7	35.8	80.8	7.8	15.1	27	5.4	0	0	0
19	84.9	32.8	664.2	3	16.7	47	4.9	0	0	0
20	89.2	27.9	70.8	6.3	15.9	35	4	0	0	0
21	86.3	27.4	97.1	5.1	9.3	44	4.5	0	0	0

Figure 5. 4: Sample Cleaned Dataset

5.4 Training the Model

After the pre-processing and evaluation of the collected dataset, training of the model immediately followed. The cleaned data set existing in a csv file was imported into a pandas Data Frame using pandas library. Figure 5.5 shows the python code used to import data and the pre-processing of the same. The labels were encoded into 0s and 1s to which is appropriate for the learning process. The features were also optimized before training using fit_transform method of the scikit-learn library as shown in the code. Thereafter, the features were split into two sets using the train_test_split method of the scikit-learn library; 80 per cent of the features were used to train the model while the remaining 20 per cent were used to test the model's performance. Figure 5.6 shows the python code used to create and train the model.

```
#import the data set
data_set = pd.read_csv('dataNew2.csv')

#preprocessing
X = data_set.iloc[:,0:8]
Y = data_set.iloc[:,8]
Y = data_set.iloc[:,8].values
for i in range(0,len(Y)):
    if Y[i] == 0.0 or Y[i] < 1.0:
        Y[i] = 0
    else:
        Y[i] = 1


X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2)
Y_labels = Y_test

Y = np_utils.to_categorical(Y)

#split to training and test
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2)
#X_one_val_test = X_test[1,0:21]


X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

Figure 5. 5: Imported data and the pre-processing

```
JavaScript output is disabled in JupyterLab
```

```
#Mode initialization and training
model = Sequential()
# input layer and the first hidden layer
model.add(Dense(units = 37, kernel_initializer = 'uniform', activation = 'relu', input_dim = 8))
# hidden layers
model.add(Dense(units = 37, kernel_initializer = 'uniform', activation = 'relu'))
model.add(Dense(units = 37, kernel_initializer = 'uniform', activation = 'relu'))
#output layer
model.add(Dense(units = 2, kernel_initializer = 'uniform', activation = 'softmax'))
#compile
model.compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics = ['accuracy'])
#train
model.fit(X_train, Y_train, batch_size =5, epochs = 400)
```

```
Epoch 1/400
413/413 [==============================] - 4s 9ms/step - loss: 0.6929 - acc: 0.5375
Epoch 2/400
413/413 [==============================] - 1s 1ms/step - loss: 0.6907 - acc: 0.5375
Epoch 3/400
413/413 [==============================] - 1s 1ms/step - loss: 0.6872 - acc: 0.5375
Epoch 4/400
413/413 [==============================] - 0s 983us/step - loss: 0.6848 - acc: 0.5375
Epoch 5/400
413/413 [==============================] - 0s 1ms/step - loss: 0.6841 - acc: 0.5375
Epoch 6/400
```

Figure 5. 6: Create and train the model

5.5 Testing the Model

The purpose of the test was to evaluate the system's compliance with the specified requirements. Twenty per cent (20%) of the features were used as test set to validate the model. This test set was passed to the trained model to be predicted and the results of the predicted compared with the actual features on the test data. Figure 5.7 depicts the testing process, prediction reports and the confusion matrix.

```
In [233]: Y_predicted = model2.predict_classes(X_test,batch_size=1,verbose=0)
          results = confusion_matrix(Y_labels, Y_predicted)
          sn.set(font_scale=1.7)#for label size
          sn.heatmap(results, cmap="Reds", annot=True,annot_kws={"size": 12})#
          print ()
          print ('Accuracy Score :',accuracy_score(Y_labels, Y_predicted))
          print ()
          print ('Prediction Report: ')
          print ()
          print (classification_report(Y_labels,Y_predicted))
          print ('Confusion Matrix :')
```

```
Accuracy Score : 0.8269230769230769

Prediction Report:

             precision    recall  f1-score   support

          0       0.85      0.85      0.85        59
          1       0.80      0.80      0.80        45

avg / total       0.83      0.83      0.83       104

Confusion Matrix :
```

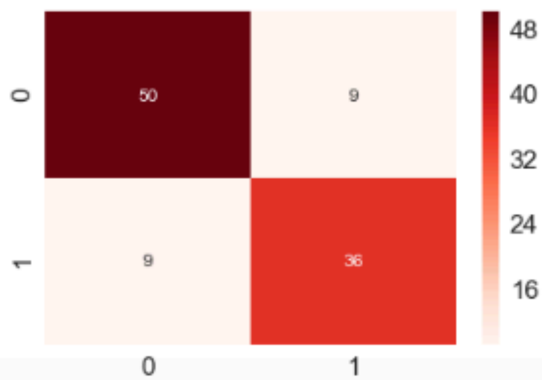

Figure 5. 7: Testing process, prediction reports and confusion matrix

The confusion matrix indicates that a total of 104 features were used to test the trained model. The comparison of the predictions against the actual labels shows that out of the 59 No labels, 50 of them were actually predicted as No and 9 were predicted as Yes, and that out of the 45 Yes labels, 36 were actually predicted as Yes and 9 were predicted as No. This prediction translates to the model accuracy score of approximately 0.83 and an average precision of 0.83 as well.

The values for the True positive, true negative, false positive and false negative were determined from the confusion matrix as demonstrated in Table 5.1.

Table 5. 1 Values from the confusion matrix

N=104	Predicted: No	Predicted: Yes
Actual: No	50	9
Actual: Yes	9	36

The metrics: accuracy, recall, precision and f1-score values are as shown in Figure 5.8. Prediction Accuracy is determined to be 82.69 per cent.

```
Accuracy Score : 0.8269230769230769

Prediction Report:

             precision    recall  f1-score   support

          0       0.85      0.85      0.85        59
          1       0.80      0.80      0.80        45

avg / total       0.83      0.83      0.83       104
```

Figure 5. 8: Metrics: accuracy, precision, recall, f1-score

A Learning curve for ANN model with varying sizes of features used was drawn to visualize the performance of the classifier as illustrated in Figure 5.9. The model improved in performance as more features were used. This implies that increasing the features in future will make the training curve and validation curve to converge.

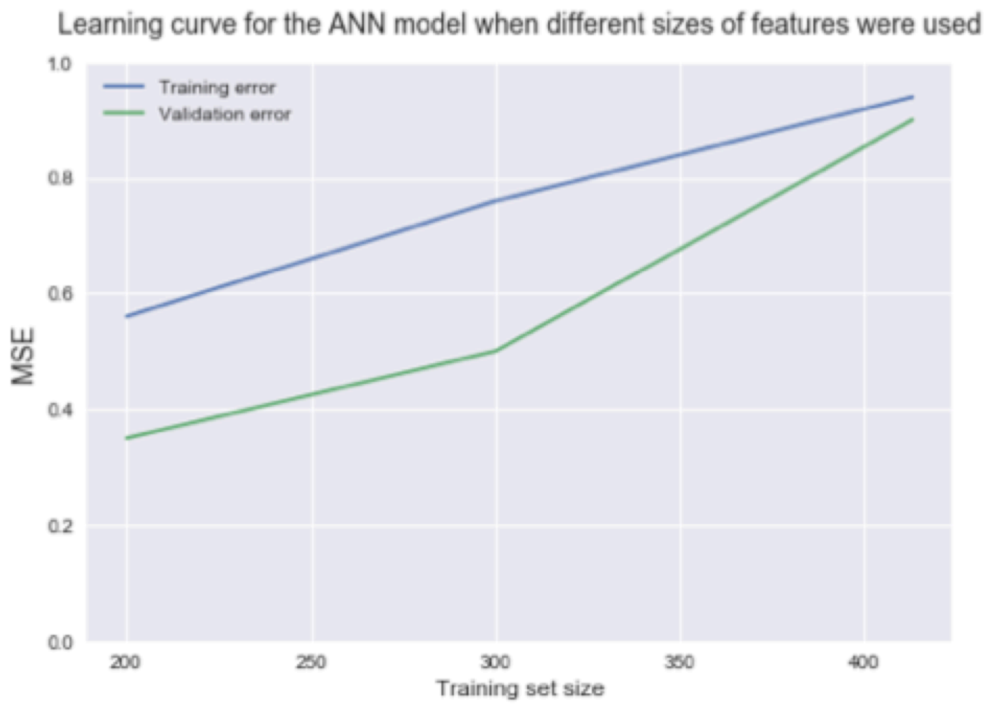

Figure 5. 9: Learning curve for the ANN model with varying sizes of features used

5.6 Using the Model in Prediction

The user is provided with a web interface to input the prediction parameters for querying the model. The model is then loaded to predict the outcome that is displayed through the same web interface that was used to input prediction parameters. Figure 5.10 shows the input/output user interphase.

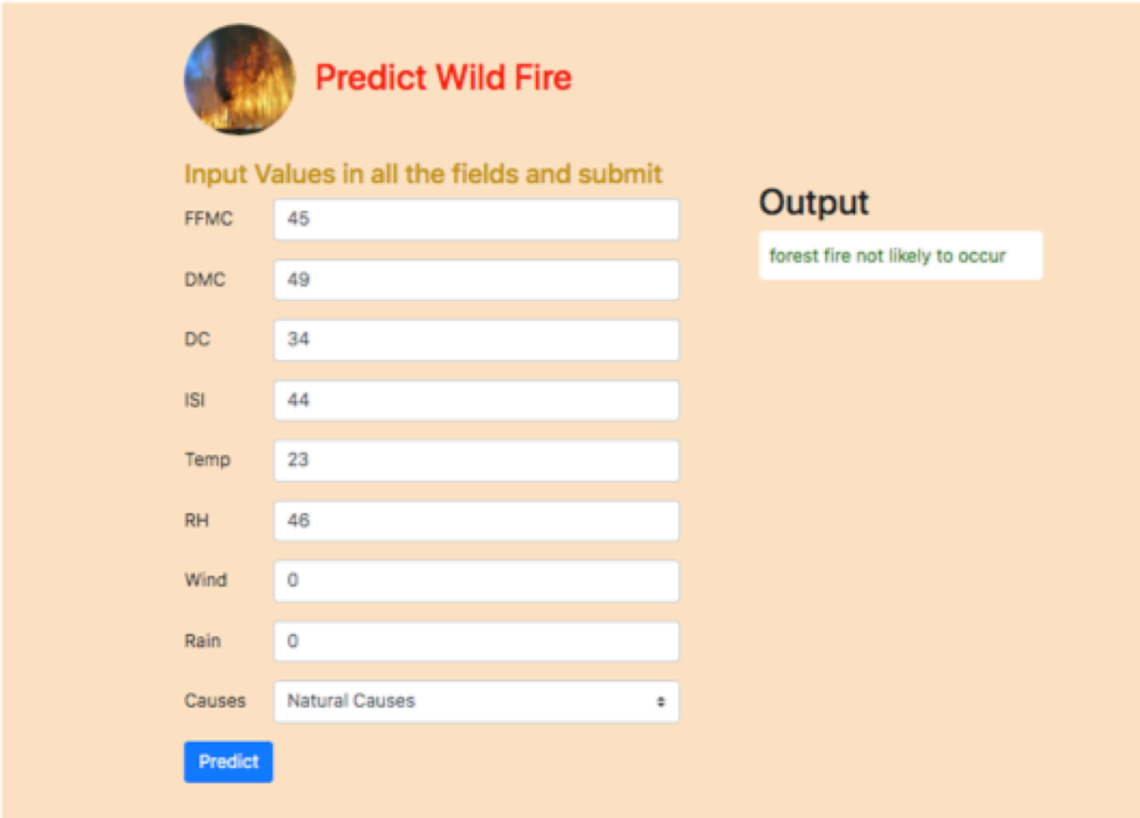

Figure 5. 10: Input/output User Interface

5.7 Implementation of Experiments

The experiments were implemented using python with the help of machine learning libraries. The scikit-learn library and keras were used for training and validating the model. The pandas and numpy libraries were used to model data before training and validation. The code for the implementation can be found in the Appendix Section (Appendix A).

Chapter 6: Discussions

6.1 Introduction

This chapter discusses the results of the developed solution compared with the set objectives and research questions in chapter one. The main objective was to develop a supervised machine-learning model to predict wild fire occurrence. ANN model was created using unigram features and tf-idf weighting with the performance tested against the test set. The model used ANN algorithm to predict wildfire occurrences.

As discussed in the problem statement, this model responds to the challenges faced by forest management stakeholders by providing a solution that accurately predicts wildfire occurrence. The study also established that 80% of wildfires are caused by human activities while past solutions have paid less attention to the effects of human factors when predicting wildfires.

6.2 Model Validation

Historical data obtained from multiple sources including National Fire Danger Rating System (NFDRS), Canada National Fire Database (CNFDB) and scientifically verified Internet sources were used in the experiments. The data was processed and fed to the ANN algorithm to give predictions based on various computations. 80% of the data were used in creating the training dataset. The remaining 20% of the data were used for validation and testing. The accuracy attained during the testing was 82.69% and it represents how well the trained model would be able to predict wildfire occurrences when presented with new records. Figure 5.8 displays the accuracy attained. The researcher conducted a number of experiments by increasing the number of features. Figure 5.9 illustrated that the model improved in performance as more features were used.

6.3 Advantages of the Developed Systems to Current Systems

The wildfire prediction system created during this research is the most significant advantage of the developed system to the currents systems. While majority of the existing solutions concentrate on localised solution, this solution attempts to offer a global approach to various stakeholders in Forest Management across the globe. On the other hand, the developed system has incorporated more factors especially the consideration of human activities in the prediction of wildfires. Majority of the existing systems do not consider fires caused as a result of human activities as they lean heavily on the meteorological factors thus creating impartial data that can lead to inaccurate predictions.

Finally, the developed system is easy to use with user-friendlier interface, which the existing solutions lack.

6.4 Shortfalls of the Research

The researcher had anticipated using more features in this study. However, the researcher had to access multiple data sets/sources in order to accumulate records with the specific attributes related to the factors necessary in the wildfire prediction. Because of the large quantity of these data sets/sources and the work needed to link the right attributes to each record, the researcher was unable to get more features.

Secondly, while considering human activities in wildfire prediction, the researcher had envisioned to have various human activities as independent variables. However, most of the data collected generalised human activities under causes just like lightning and electrical sparks. This therefore forced the researcher to consider human activities as a dependable variable.

Chapter 7: Conclusions, Recommendations and Future Work

7.1 Conclusions

The study was focused on developing a machine-learning algorithm for predicting wildfire occurrence. There was need to review relevant literature to determine the challenges encountered by the stakeholders in the prediction of wildfires. Equally, there was need to review literature to understand the existing techniques and models for wildfire prediction.

The main objective of the study was to develop a supervised machine-learning model to predict wildfire occurrence. To achieve this, a number of processes were undertaken. Data was collected from credible climatological sources that included both meteorological sources as well as wildfire databases focusing on content dated between 2000 and 2020; all the major components contributing to the ignition of wildfires were put into consideration. The data was preprocessed and split into training and test sets. The training set was used to train the ANN model using set features with tf-idf weighting. The testing set was used to test the model, which achieved an accuracy of 82.69 percent. The study revealed the relationship between climatological factors, human activities and fire counts in the prediction of wildfire occurrence.

7.2 Recommendations

This work demonstrates that ANN model can be used to predict wildfire occurrences instead of the use of human beings and analogue technologies as done by the Kenya Forest Services and it will be helpful in the prevention and rescue of forest fires. Forest management stakeholders will be able to take effectual and applicable measures as per the system's predictions. To this point, the researcher recommends:

i. Use of large data set for better prediction results. A total number of 20000 data volumes were collected but only 414 (80 per cent) records were used to train the model and the remaining 103 (20 per cent) used to test and validate the model. There is need to increase the size of the training data to improve the performance.

7.3 Future Work

Human activities have a huge impact on the occurrence of forest fires. This study takes into account the general effects of human activities in the wildfire occurrence. However, more research needs to be done on the specific human activities and analyse the significance of each to the occurrence of wildfires.

References

Aseko, J. (2018). Application of Forest Fire Detection Index in Forest Fires (Thesis, University of Nairobi). Retrieved from http://erepository.uonbi.ac.ke/bitstream/handle/11295/104217/FINAL%20MSC%20PROJECT-REVISED%20%281%292.pdf?sequence=1&isAllowed=y

Bonsor, K. (2001). How Wildfires Work. Retrieved from https://science.howstuffworks.com/nature/natural-disasters/wildfire1.htm

Bracmort, K. (2013). Forest Fire/Wildfire Protection. Retrieved February 22, 2020, from https://pdfs.semanticscholar.org/8f0a/d33b3c28b62b76ac2626de92eb017dc656ce.pdf?_ga=2.161230840.478705711.1582098184-763335520.1582098184

Camp, P., & Dorflinger, A. (2018). How Humans Shape the Wildfire Peril. Retrieved February 15, 2020, from https://www.air-worldwide.com/blog/posts/2018/2/how-humans-shape-the-wildfire-peril/

Chakrabarty, H., & Bhattacharya, S. (2015). Application of K-Nearest Neighbor Technique to Predict Severe Thunderstorms. *International Journal of Computer Applications*, *110*(10), 1–4.

Chas-Amil, M. L., Prestemon, J. P., McClean, C. J., & Touza, J. (2015). Human-ignited wildfire patterns and responses to policy shifts. Retrieved February 20, 2020, from https://www.fs.usda.gov/treesearch/pubs/48355

Cohen, J. D., & Deeming, J. E. (1985). The National Fire-Danger Rating System: basic equations. Retrieved February 22, 2020, from https://www.fs.fed.us/psw/publications/documents/psw_gtr082/psw_gtr082.pdf

Dale, V.H., Joyce, L.A., McNulty, S., Neilson, R.P., Ayres, M.P., Flannigan, M.D., Hanson, P.J., Irland, L.C., Lugo, A.E., Peterson, C.J., Simberloff, D., Swanson, F.J., Stocks, B.J., and Wotton, B.M., (2001). *Climate change and forest disturbances*

Davies, S. J. & L. Unam (1999). Smoke-haze from the 1997 Indonesian forest fires: effects on pollution levels, local climate, atmospheric CO2 concentrations, and tree photosynthesis. *Forest Ecology and Management*

Dey, A., Bhoumik, D., & Dey, K. (2016). Automatic Detection of Whitefly Pest using Statistical Feature Extraction and Image Classification Methods. International Research Journal of Engineering and Technology,3(9), 950-959. Retrieved February 21, 2020

Doyle, K., Deacon, B., & Locke, S. (2017, December 12). Fire danger system being reviewed to cope with new extremes. Retrieved from https://www.abc.net.au/news/2017-12-13/bushfire-danger-rating-system-trialled-summer/9203446

Ganteaume, A., Camia, A., Jappiot, M., San-Miguel-Ayanz, J., Long-Fournel, M., & Lampin, C. (2012). A Review of the Main Driving Factors of Forest Fire Ignition Over Europe. *Environmental management.*

Guo, F., Zhang, L., Jin, S., Tigabu, M., Su, Z., & Wang, W. (2016, October 25). Modeling Anthropogenic Fire Occurrence in the Boreal Forest of China Using Logistic Regression and Random Forests. Retrieved from https://www.mdpi.com/1999-4907/7/11/250

Harrison, O. (2019, July 14). Machine Learning Basics with the K-Nearest Neighbors Algorithm. Retrieved from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Henry, M. C., McCarty, J., & Maingi, J. (2020, January 27). Why Mount Kenya is burning and what can be done to stop the fires. Retrieved February 15, 2020, from https://theconversation.com/why-mount-kenya-is-burning-and-what-can-be-done-to-stop-the-fires-112842

Kapaya, L. (2017). Rapid Application Development. Retrieved from https://www.slideshare.net/LombeKapaya/rapid-application-development-72725864

Kothari, C. (2003). *Research Methodology: Methods & Techniques*. Retrieved from http://books.google.co.ke/books

Krebs P., Pezzatti G.B., Mazzoleni S., Talbot, L.M., & Conedera, M. (2010). Fire regime: history and definition of a key concept in disturbance ecology. *Theory in Biosciences*

Liang, H., Zhang, M., & Wang, H. (2019). A Neural Network Model for Wildfire Scale Prediction Using Meteorological Factors. *IEEE Access, 7*, 176746-176755. doi:10.1109/access.2019.2957837

Mann, M. L., Batllori, E., Moritz, M. A., Waller, E. K., Berck, P., Flint, A. L., … Dolfi, E. (2016). Incorporating Anthropogenic Influences into Fire Probability Models: Effects of Human Activity and Climate Change on Fire Activity in California. Retrieved February 15, 2020, from https://journals.plos.org/plosone/article?id=info:doi/10.1371/journal.pone.0153589

Manning, C., Raghavan, P., & Schutze, H. (2009). An Introduction to Information Retrieval. Cambridge: Cambridge University Press.

Marchal, J., Cumming, S. G., & Mcintire, E. J. B. (2016). Exploiting Poisson additivity to predict fire frequency from maps of fire weather and land cover in boreal forests of Québec, Canada. *Ecography*, *40*(1), 200–209. doi: 10.1111/ecog.01849

Mickleburgh, P. (2009). How topography effects fire movement and impacts stand structure in a Boreal forest.

Moore, P., Hardesty, J., Kelleher, S., Maginnis, S., & Myers, R. (2003). Retrieved February 15, 2020, from http://www.fao.org/3/xii/0829-b3.htm

Naganathan, H., Seshasayee, S., Kim, J., Chou, J., & Chong, W. (2016). Wildfire Predictions: Determining Reliable Models using Fused Dataset. *Global Journal of Computer Science and Technology*. (16) 35-46

Nyaga, M. (2019).A Mobile-based Image Recognition System for Identifying Bird Species in Kenya (Thesis, Strathmore University). Retrieved from https://su-plus.strathmore.edu/bitstream/handle/11071/6610/Gideon.pdf?sequence=1&isAllowed=y

Patel, K. (2018). Six trends to know about fire season in the western U.S. – Climate Change: Vital Signs of the Planet. Retrieved from https://climate.nasa.gov/blog/2830/six-trends-to-know-about-fire-season-in-the-western-us/

Plucinski, M. P., McCaw, W. L., Gould, J. S., & Wotton, B. M. (2014, April 9). Predicting the number of daily human-caused bushfires to assist suppression planning in south-west Western Australia. Retrieved from https://www.publish.csiro.au/wf/WF13090

Pupale, R. (2019). Support Vector Machines (SVM) - An Overview. Retrieved from https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

Reduce Wildfire Risk - Washington Forest Protection Association. (n.d.). Retrieved from http://www.wfpa.org/sustainable-forestry/reduce-wildfire-risk/

Rishickesh, R., Kumar, R., Shahina, A., & Khan, A. N. (2019). Identification of Duplication in Questions Posed on Knowledge Sharing Platform Quora using Machine Learning Techniques. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *8*(12), 2444–2451. doi: 10.35940/ijitee.L3017.1081219

Robert, C. (2014). Machine Learning, a Probabilistic Perspective. Chance, 27(2), 62-63.

Schlobohm, P., & Brain, J. (2002). Gaining an Understanding of the National Fire Danger Rating System. Retrieved February 20, 2020, from https://www.nwcg.gov/sites/default/files/products/pms932.pdf

Siva, F. (2019). Smart fertilizer recommendation through NPK analysis using Artificial Neural Networks (Thesis, Strathmore University). Retrieved from http://su-plus.strathmore.edu/handle/11071/6702

Sukumar, K. K., Joevivek, V. K., Hemalatha, T. K., & Soman, P. K. (2009). Prediction of Forest Fire Using SVM – A Novel Approach. In *10th ESRI India User Conference 2009.* Retrieved from https://www.esri.in/~/media/esri-india/files/pdfs/events/uc2009/papers/DM3.pdf

Sun, P., & Zhang, Y. L. (2018). A probabilistic method predicting forest fire occurrence combining firebrands and the weather fuel complex in the northern part of the Daxinganling region, China. *Forests*, 9(7), 428. https://doi.org/10.3390/f9070428

Tobin, M. (2013). Retrieved from http://ecowest.org/2013/06/04/wildfire-ignition-trends-humans-versus-lightning/

Toledo-Castro, J., Caballero-Gil, P., Rodríguez-Pérez, N., Santos-González, I., Hernández-Goya, C., & Aguasca-Colomo, R. (2018, December 2). Forest Fire Prevention, Detection, and Fighting Based on Fuzzy Logic and Wireless Sensor Networks. Retrieved from https://www.hindawi.com/journals/complexity/2018/1639715/

Topography's effect on Fire Behavior. (n.d.). Retrieved from http://www.auburn.edu/academic/forestry_wildlife/fire/topos_effect.htm

Wolters, K. (2019). WOLTERS KLUWER ENVIRONMENTAL POLICY. Retrieved from https://wolterskluwer.com/binaries/content/assets/wk/pdf/corporategovernance/policies-and-articles/190211-wolters-kluwer-environmental-policy.pdf

Younes, O. S., Mousannif, H., & Moatassime, H. (2019). Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire Safety Journal*. 104. 10.1016/j.firesaf.2019.01.006.

APPENDIX A

Full Code

```
import keras

from keras import backend as K

from keras.models import Sequential #For Initializing ANN

from keras.layers import Activation

from keras.layers.core import Dense #For Layers of ANN

from keras.optimizers import Adam

from keras.metrics import categorical_crossentropy

import numpy as np

from random import randint

from sklearn.preprocessing import MinMaxScaler,LabelEncoder,OneHotEncoder,StandardScaler

from sklearn.metrics import confusion_matrix

import itertools

import matplotlib.pyplot as plt

from keras.utils import np_utils

from sklearn.model_selection import train_test_split

from sklearn.externals import joblib

from sklearn.metrics import accuracy_score

from sklearn.metrics import classification_report

label_encoder_x_1 = LabelEncoder()

sc = StandardScaler()

import seaborn as sn

#import the data set

#data_set = pd.read_csv('dataNew2.csv')
```

```
data_set = pd.read_csv('dataNew.csv')
#preprocessing
X = data_set.iloc[:,0:8]
Y = data_set.iloc[:,8]
Y = data_set.iloc[:,8].values
for i in range(0,len(Y)):
  if Y[i] == 0.0 or Y[i] < 1.0:
     Y[i] = 0
  else:
     Y[i] = 1
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2)
Y_labels = Y_test
Y = np_utils.to_categorical(Y)
#split to training and test
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2)
#X_one_val_test = X_test[1,0:21]
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
#Mode initialization and training
model = Sequential()
# input layer and the first hidden layer
model.add(Dense(units = 37, kernel_initializer = 'uniform', activation = 'relu', input_dim = 8))
# hidden layers
model.add(Dense(units = 37, kernel_initializer = 'uniform', activation = 'relu'))
model.add(Dense(units = 37, kernel_initializer = 'uniform', activation = 'relu'))
```

```
#output layer

model.add(Dense(units = 2, kernel_initializer = 'uniform', activation = 'softmax'))

#compile

model.compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics = ['accuracy'])

#train

model.fit(X_train, Y_train, batch_size =5, epochs = 400)

#export trained model and test

joblib.dump(model,'/Users/admin/Desktop/ForestFirePrediction/ForestFireTraineModel')

#import model

model2 = Sequential()

model2 = joblib.load('/Users/admin/Desktop/ForestFirePrediction/ForestFireTraineModel')

#Y_predicted = model2.predict_classes(X_test,batch_size=1,verbose=0)

#prediction and validation

Y_predicted = model2.predict_classes(X_test,batch_size=1,verbose=0)

results = confusion_matrix(Y_labels, Y_predicted)

sn.set(font_scale=1.7)#for label size

sn.heatmap(results, cmap="Reds", annot=True,annot_kws={"size": 12})#

print ()

print ('Accuracy Score :',accuracy_score(Y_labels, Y_predicted))

print ()

print ('Prediction Report: ')

print ()

print (classification_report(Y_labels,Y_predicted))

print ('Confusion Matrix :')

#plot learning as per the iterations done with different faeture size
```

```
import matplotlib.pyplot as plt

%matplotlib inline

train_sizes = [200,300,413]

train_scores_mean= [0.56,0.76,0.94]

validation_scores_mean = [0.35,0.50,0.90]

plt.style.use('seaborn')

plt.plot(train_sizes, train_scores_mean, label = 'Training error')

plt.plot(train_sizes, validation_scores_mean, label = 'Validation error')

plt.ylabel('MSE', fontsize = 14)

plt.xlabel('Training set size', fontsize = 12)

plt.title('Learning curve for the ANN model when different sizes of features were used',

        fontsize = 14, y = 1.03)

plt.legend()

plt.ylim(0,1)

#Engine to consume above model and present results to user

import numpy as np

from flask import Flask, render_template, request

from keras.models import Sequential

from sklearn.externals import joblib

model2 = Sequential()

model2 = joblib.load( 'ForestFireTrainedModel' )

app = Flask( __name__ )

@app.route( "/", methods=["POST", "GET"] )

def home():

    if request.method == "POST":
```

```
        data = []
        ffmc = request.form["ffmc"]
        dmc = request.form["dmc"]
        dc = request.form["dc"]
        isi = request.form["isi"]
        temp = request.form["temp"]
        rh = request.form["rh"]
        wind = request.form["wind"]
        rain = request.form["rain"]
        data.append( ffmc )
        data.append( dmc )
        data.append( dc )
        data.append( isi )
        data.append( temp )
        data.append( rh )
        data.append( wind )
        data.append( rain )
        result = get_result( data )
        return render_template( "index.html", result=result, ffmc=ffmc,
                        dmc=dmc, dc=dc, isi=isi, temp=temp, rh=rh, wind=wind, rain=rain )
    else:
        return render_template( "index.html" )
def get_result(data):
    npa = np.array( [data] )
    outcome = model2.predict_classes( npa, batch_size=1, verbose=0 )
```

```
    return outcome[0]

if __name__ == "__main__":

    app.run( debug=True )

#front code for user interface

{% extends "base.html" %}

{% block title %}home{% endblock %}

{% block content %}

<div class="row">

    <div class="col-lg-2">

    </div>

    <div class="col-lg-5" style="margin-top: 20px;">

        <h2 style="color: red;margin-top: 20px;margin-bottom: 20px;">

            <img style="margin-right: 10px;" height="100" width="100" class="rounded-circle"

                 src="../static/images/wf.png">

            Predict Wild Fire </h2>

        <h4 style="color: #c69500" style="margin-bottom: 40px;">Input Values in all the fields and
submit </h4>

        <form action="#" method="post">

            <div class="form-group row">

                <label for="ffmc" class="col-sm-2 col-form-label">FFMC</label>

                <div class="col-sm-10">

                    <input type="text" class="form-control" id="ffmc" name="ffmc" placeholder="FFMC"
required

                           value="{{ffmc}}">

                </div>

            </div>
```

```
<div class="form-group row">
  <label for="dmc" class="col-sm-2 col-form-label">DMC</label>
  <div class="col-sm-10">
    <input type="text" class="form-control" id="dmc" name="dmc" placeholder="DMC"
required
        value="{{dmc}}">
  </div>
</div>
<div class="form-group row">
  <label for="dc" class="col-sm-2 col-form-label">DC</label>
  <div class="col-sm-10">
    <input type="text" class="form-control" id="dc" name="dc" placeholder="DC"
required value="{{dc}}">
  </div>
</div>
<div class="form-group row">
  <label for="isi" class="col-sm-2 col-form-label">ISI</label>
  <div class="col-sm-10">
    <input type="text" class="form-control" id="isi" name="isi" placeholder="ISI"
required
        value="{{isi}}">
  </div>
</div>
<div class="form-group row">
  <label for="temp" class="col-sm-2 col-form-label">Temp</label>
  <div class="col-sm-10">
```

```
                <input type="text" class="form-control" id="temp" name="temp"
placeholder="Temp" required
                    value="{{temp}}">
            </div>
        </div>
        <div class="form-group row">
            <label for="rh" class="col-sm-2 col-form-label">RH</label>
            <div class="col-sm-10">
                <input type="text" class="form-control" id="rh" name="rh" placeholder="RH"
required value="{{rh}}">
            </div>
        </div>
        <div class="form-group row">
            <label for="wind" class="col-sm-2 col-form-label">Wind</label>
            <div class="col-sm-10">
                <input type="text" class="form-control" id="wind" name="wind"
placeholder="Wind" required
                    value="{{wind}}">
            </div>
        </div>
        <div class="form-group row">
            <label for="rain" class="col-sm-2 col-form-label">Rain</label>
            <div class="col-sm-10">
                <input type="text" class="form-control" id="rain" name="rain" placeholder="Rain"
required
                    value="{{rain}}">
            </div>
```

```
          </div>

          <div class="form-group row">

            <label for="causes" class="col-sm-2 col-form-label">Causes</label>

            <div class="col-sm-10">

              <select class="custom-select" id="causes" name="causes">

                <option value="1">Natural Causes</option>

                <option value="2">Human Activity</option>

              </select>

            </div>

          </div>

          <div class="form-group row">

            <div class="col-sm-10">

              <input type="submit" class="btn btn-primary" value="Predict">

            </div>

          </div>

        </form>

      </div>

      <div class="col-lg-4" style="margin-top: 180px;margin-left: 40px">

        <h2>Output</h2>

        {% if result == 1 %}

        <p style="color: red;background-color:white;padding: 10px;border-radius: 5px;font-weight:
bolder">forest fire

          most likely to occur</p>

        {% elif result == 0 %}

        <p style="color: green;background-color:white;padding: 10px;border-radius: 5px;font-
weight: bolder">forest fire
```

```
        not likely to occur</p>

      {% endif %}

   </div>

</div>

{% endblock %}
```

APPENDIX B

Originality Report

Turnitin Summary Report