Wild fire prediction using machine learning models

Nadir Khan
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
tp042831@mail.apu.edu.my

Nowshath K Batcha
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
nowshath.kb@apu.edu.my

***Abstract*— Wildfires can be very destructive and can cause a great loss to human life and property. The United States have witnessed an estimated loss of 13.7 billion dollars' worth of economic losses from 2002 till 2011 due to wild Fires. Moreover countries of Southern Europe (i.e. Italy, Spain, and Portugal) have witnessed increased wildfires during last decade. Deforestation poses a major risk to the environment and wild Fires causes an unprecedented increase in Deforestation by increasing the burned area along with burning of trees, plants which is very vital to maintain a healthy environment. Moreover deforestation causes flooding and mud sliding thereby have an adverse impact on the ecosystem. Hence this study focusses on building wild fire prediction model based on machine learning that will facilitate to take precautionary measures in order to minimize the losses to human life and property.**

Keywords—recognition, convolution, neural

I. INTRODUCTION

Forest fires incidents have increased dramatically over the years. Uncontrolled Wildfires can be very destructive and can cause a great loss to human life and property [1]. Destructive wildfire can result in vast damage especially where it is closed to cities or town where human population exists. Among major disaster that can cause heavy damage to economy of any country other than Wildfires are Floods, tsunamis, hurricanes, earthquakes. Forest Fires are generally referred to a situation where fires under the influence of climatic and topographical conditions are more likely to cause wildfire destruction and produce accidents. Any country victim of the latter can cause substantial damage to human population and infrastructure for any country. Natural fires have long played a significant role in terrestrial ecosystem. Particularly the countries of Southern Europe (i.e. Italy, Spain, and Portugal) have witnessed increased wildfires during last decade. From 2002 till 2011, wildfires in US accounted for a total of 13.7 billion dollars of economic losses. This increase in economic losses is 6.9 billion dollars more than the previous decade [1].

Moreover a total of 13 Fire Fighters have lost their life while tackling the wildfires which is recorded as the largest loss since 2013.Wildfire risk have increased for human population by reaching in close vicinity to towns and major cities. An estimate revealed that about 32% of units including buildings and apartments and 10% of all lands in the United States are located near the wildlife which is closed to the occurrence of wildfire incidents. Wild lands which consist of housing units, apartments and buildings and the area where wildfire occurs is known as wild land urban interface abbreviated as WUI.

Human population located in wild land urban interface are highly vulnerable to fire irrespective of size of fire or any vegetation type. As per the statistics revealed by United States of Agriculture 1.6 billion dollars are spent annually by the state forestry agencies on wildfire protection and preventive measures. Forest Fires majority of the time occurs in remote areas which are hard to access and usually have no proper routes for transportation of fire equipment. Fire whirl is another form of Wildfire. Fire whirl is a situation which under the influence of certain conditions acquires a vertical vortices and generates a flame in a vertical direction that is labelled as a fire tornado. It consist of massive amount of energy accompanied with erratic movement making it as one of dangerous behavior associated with fire behavior. The erratic movement of fire whirls enable fire to transport fire to areas far away quickly and this behavior allows to the Fire to spread itself quickly by spotting due to embers that are lifted in the core to a greater height and transported far away from fire perimeter by wind. Research have been conducted to find the exact causes of fire whirls which yields maximum damages along with their influence on wildfire. Some studies have revealed the significant role of climate including topography and fuel as the significant factors behind the generation of fire whirls and help to predict its behavior. Another Notable forest known as Amazonian forest located in Brazil account for 35% of worlds tropical forest carbon and produces largest emission from forest fires.

Amazonian forest in Brazil have also been described as a success story for a major reduction in deforestation based on effective public policies. Humans and Lightening's account as one of major reasons for wildfire incidents. Areas or Regions having summer droughts are highly vulnerable to wildfire. Lightning can ignite forest fires up to 40% or more. Lightings occur everywhere in the world and not restricted to any specific area or region. Lightning fires are often difficult to combat by fire brigades and may lead to deforestation on a massive scale. Lightning usually occurs at night time and may be facilitated by weather conditions which are appropriate for ignition. Majority of the lightning occurs in regions with large forested areas like Canada, Europe, US and some regions in Europe. The impact of lightening or some characteristic of lightning that causes wildfire are often hard to find out because of the accessibility and the time the lightening occurs.

In USA and Australia approximately 95% of all wildfires are tackled immediately on emergency basis in order to minimize the intensity of fire while around 4% of the fire erupted requires extended operations and may often require one or 2 minimum burning periods. About 1% of wildfire

transforms into a wide scale thus requires the setting up of organized Command Control System. The purpose behind setting up of Organized Command Control System is to monitor the planning, logistical and operation leadership for the whole ongoing operation to control the wildfire which may result in large scale of burned area.

As population in the Western United States increases, the alarming point to consider here is that more human population along with housing units are approaching to the wild land areas. As earlier discussed the wild land urban interface where people live are highly vulnerable to the exposure of wildfire. Continued increase in development and human population, damages as a result of wildfire are now getting more costly.

This all clearly necessitates the demand of any security mechanism that can foresee the occurrence of the wild fire that can serve to society.

II. Literature Review

Hotter temperatures results in drying of fuels required for fire ignition as the temperature increases in spring and summer season. Drying of fire fuels results in downed wood on the earth in the forest and standing trees [2]. Fire are more likely to start in hotter and drier regions in case Lightning strikes or human activity. Western Regions have witnessed an increase of 2.1 F as compared to last 45 years. Table 1 below shows the rank of 11 States based on spring summer temperature since 1979's.New Mexico, Arizona and Colorado have witnessed higher increase in temperature since 1970's.

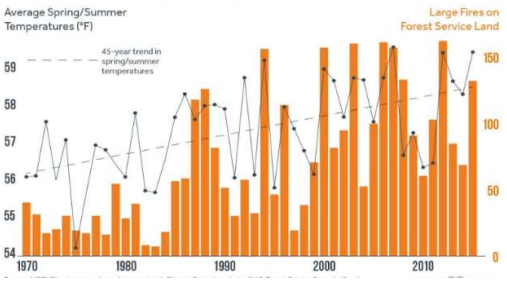

Fig. 1. Temperature Trends since 1970 [2]

Hotter Temperatures result in early meltdown of snow causing a decrease in availability of water during the hotter season of the year. An increase in Temperature of 2 F have been recorded across the Western regions as compared to the last 45 years accompanied with higher increase in Temperature have been recorded in Arizona, New Mexico and California. Most of climate models remain uniform when assessing temperatures and Precipitations. But on the contrary in Alaska amount of warming entirely depends on emissions. Temperature increase across Western States lead to earlier melting of snow, a significant decrease in precipitation level during fire season which facilitates ignition of wildfire effectively.

Climate central examined the historical climate data along with trend for Climate in order to assess the wild fire risk in the coming decades. The likelihood for Wildfire across Western States might increase till 2050.The analysis of these likelihood are based on Ketch –Byram Drought Index (KBDI) which basically evaluates the dryness of the forest floor. The Scale of KBDI is based on a range of 0-800 where the low numbers reveal that the moisture is high (and fuel is less likely to burn) while the high numbers indicate severe drought and higher likelihood of fires. Number of days with KBDI above would increase significantly in 10 States that fall in Western Region if the greenhouse gas emissions continue to increase.

Fire management in Australia's Heritage listed Kakadu National park is based on the information that consist of the science of how to monitor and manage the natural resources Kakadu National Park in Australia is inhabited by the Aborigines (McGregor et al., 2010). Aboriginal society consist of individual having their own way of lifestyle and adaptation to the environment. Based on varying style of living every individual that belongs to Aboriginal society pursues his own way to managing his land. Fire management within the Kakadu seems to be a major issue in the Northern Australia. The burning of wetlands along with the burning of surrounding of Savanna Woodlands has drawn a considerable amount of focus from the scientific community. The Ramsar wetlands which lies adjacent to the floodplains to Kakadu's major river system is extremely vital for Aboriginal people who resides there [3].

Floodplains consist of swamps, water grasslands and sedge lands. The most recent changes that have been witnessed by Kakadu's wetlands was due to presence of Asian Water Buffalo also known as Bubalus bubalis. The Asian Water Buffalo presence have result in decrease of biomass of the grass. This decrease in the biomass of the grass directly affected the fuel for fires. Moreover it resulted in mixing of salt water into the fresh water swamps. Burning of the flood plain usually occurs at the end of the dry season which usually starts in September. The dry season can facilitate the ignition of fires .To prevent Wild fires flood margins have already burned to restrict the wild fire from spreading further into the savanna woodlands.

Deliberate fires occurs because of hunting, early burning, grazing, cooking or dumping. Other reasons for deliberate fires may include cooking, smoking cigarettes, religious and traditional festivals. Accidental fires on the other hand mainly occurs due to the movement of locomotives on the Bulawayo Victoria lines, lightning or combustion during the dry season [4]. Poor farmers on the account of possessing limited resources mainly use mechanical methods to prepare their land for cultivation and may sometimes use fire for land preparation. Fires erupted during land preparation may spread out and then turn into wild fires depending on the intensity of the fire.

Moreover one of the significant ecological effect of burning is the chances of further burning in the coming years as dead trees vanish exposing the forest to more sunlight thereby increasing the chances of fuel fire. Wild fires that occur in Tropical forest result in decrease of woody plant richness by two thirds. Fire also damages the seed bank, individual seeds and plants. Occurrence of wild fires give rise to fire tolerant species i.e. trees that have thicker insulating bark. Fires also upsets the ecosystems by reducing the fruit eating birds, amphibians and reptiles.

Wildfires have an adverse on the human health and environment no matter how far it is from human population. National Climate Assessment has revealed the adverse impact of Wildfires on human health. It further states that smoke arising from Wildfires can give rise to respiratory and cardiovascular issues for human health. Most important it

transforms the quality of air we breathe in to a bad quality as per the findings of Climate Central analysis. The Western Region of the US had bad air quality during the occurrence of Wildfire.

Moreover as result of Wildfires that occurred in certain regions in China the local air quality became bad similar to that of Beijing, the capital of China having bad air quality. Another such example is Houston where the smoke arising from Wildfires in Alaska and Canada caused the ozone levels harmful. A Climate analysis of overall 45 years from records of US Forest Service Records for the Western region reveal that number of large fires in the western regions is increasing dramatically. Moreover the burned area as result of Wildfires are also increasing exponentially.

III. Problem Context

The occurrence of wild fire has an adverse effect on the economy of any country or region. Forest fires are another form of natural disaster which causes an unprecedented loss to the economy in terms of social and environmental context. In some cases forest fire can endanger the life of those individuals who are in charge of suppressing it.

Forest fires under the influence of weather and topography conditions are more likely to cause widespread destruction. Humans and lightening account for most of fire ignitions around the World. Especially in regions under the influence of drought combined lighting can increase forest fires up to 40%.Lightings occur almost everywhere and are not confined to any particular area.

Forest fires are often located in remote areas and very difficult to combat it with help of fire brigades. Lightning at night time makes firefighting very difficult to tackle and increases burned area and severity. Recently in December 2017 a total of 200,000 people have been evacuated in Californian due to raging Wild fire leaving 40 people dead and causing $85 billion dollars loss to the economy. In this case study having a sample dataset of US Wildfires we will predict the occurrence of Wildfires.

IV. Methodology

The Methodology selection adopted for this study was CRISP-DM.CRISP-DM stands for Cross industry standard process for data mining. CRISP-DM basically defines the set of tasks, output generated from these tasks along with terminology and mining problem. Among the industrial pioneers who adopted CRISP-DM standard for carrying out data mining were Daimler Chrysler, SPSS and NCR (Anand et al., 2007)..CRISP –DM basically divided the knowledge discovery process into six phases shown in Fig 2. CRISP –DM is neither a Waterfall model nor it is similar to rapid prototyping model.

The major guidelines for adopting CRISP-DM is to carrying out the each task carefully before proceeding to the next step. Backtracking to the previous task is also possible for knowledge discovery (Anand et al., 2007). For instance during modelling phase it is discovered that perhaps some additional preprocessing should have done prior to generating the models. The six phases of CRISP-DM has been 3 and 7 Generic task by the CRSIP-DM consortium. A generic task merely divides the top level phases into sub tasks to be carried in this phase.

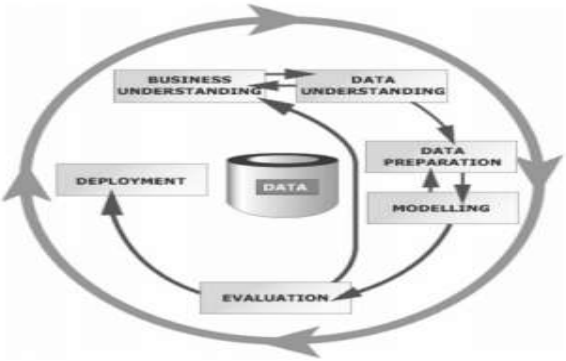

Fig. 2. Six Phases for CRISP-DM Methodology [5]

The machine learning techniques selected for developing the model is explained.

Decision tree are powerful data mining techniques. Decision Tree is one of the fundamental techniques used by regression and classification methods represented by Tree structure [6]. Regression or classification results usually represents by the leaves in the decision tree. A decision tree is a visual representation of if then form of rule based on determining results. Similar to other techniques it can perform multiple variable analysis. Decision tree mainly works in by splitting the data into segments also called branches. The splitting of the data eventually produces an inverted decision tree with a root node at the top of the tree. Decision tree algorithm basically works on features and attributes in a dataset using small space and taking advantage of fast classification speed.

Classification is another technique used extensively in field of data mining in order to forecast a categorical value by generating a model based on target variable which is a part of dataset .It is a type of supervised learning. The target variable can contain numerical or categorical value. Classification algorithm is meant to understand the hidden pattern of the database structure for prediction [7]. There are other Classification techniques like Bayesian networks, lazy classifier and rule based classifier.

Logistic regression has widely been used in different applications such as generating model for dose response data or purchase choice data. Logistic regression model requires category response variable and at least one effect variable or interaction term. Logistic regression analysis basically analyzes that association between a categorical dependent variable and independent variables. Logistic Regression is frequently used when dependent variable has 2 values such as 0 or 1. Another type of logistic regression known as Multinomial logistic regression when the Dependent variable has three or more unique values such as Married, Single or divorced or widowed. Logistic regression shares the similarity with discriminant analysis because both analyzes categorical response variables.

Random forest is a type of learning technique that consist of many classification trees which are then combined to compute a classification tree (Al-Abadi & Shahid, 2016).

Random forest is ensemble of learning techniques. Learning techniques of type ensemble generates many classifiers and combine their results. Ensemble learning

techniques can be split into 2 categories which as bagging and boosting.

Dataset has been acquired from kaggle.com which contains a repository for dataset. The dataset is mainly used for research and academic purposes. The dataset for our research contains data related to the occurrence of Wild fires in US. The dataset was extracted as SQLite file from kaggle.com which is labelled as FPA_FOD_20170508.sqlite. The file was opened with SQLite Studio. SQLite Studio is a basically a Graphical User Interface for writing and debugging SQL queries using SQLite database.FPA_FOD_20170508.sqlite file contain several database tables.

V. EXPERIMENTATION

Performance Measurement consist of Statistical methods which are primarily used to assess the overall performance of model using statistical measures .It is of considerable importance in determining the quality of the system and can also be used to optimize the system depending on the requirements for which it is designed. There are several techniques for assessing performance measures of a model but in this research we will take into account Accuracy, RMSE, Precision and Recall for evaluating the model.

Performance Evaluation of Model is very important as all our analysis depends on it. It is the final outcome of our prediction model including all the phases. All the details about training and Validation of the dataset have already been discussed. Figures below shows the comparison of RMSE, Accuracy, Precision and Recall for all the Four Techniques. Bar chart below shows the accuracy of all the four techniques for model evaluation.

Technique	Accuracy	Precision	Sensitivity	RMSE
Decision Tree	0.92	0.92	0.92	0.1
Classification	0.95	0.95	0.95	0.1
Logistic Regression	0.82	0.81	0.82	2.9
Random Forest	0.92	0.92	0.92	1.3

Fig. 3. Model Output

As we can see from Figure 4 below regarding comparison of Accuracy of the Techniques. Among all the four techniques applied for generating the model, Classification (KNN) yields the highest Accuracy (95%).Decision Tree and Random Forest yields the same Accuracy (92%) while Logistic Regression yields the lowest Accuracy (82%). Fig 4. shows the Comparison of Accuracy of all the four techniques used to build the model.

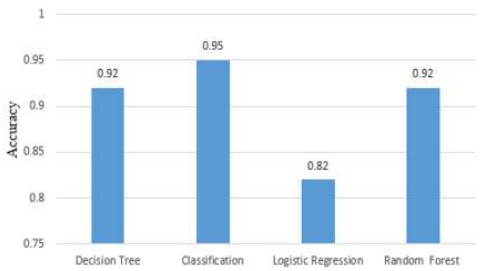

Fig. 4. Comparison of accuracies

As we can see from Fig 5, regarding comparison of Precision of the Techniques. Among all the four techniques applied for generating the model. Classification (KNN) yields the highest Precision (95%).Decision Tree and Random Forest yields the same Precision (92%) while Logistic Regression yields the lowest Precision (81%). Figure below shows the Comparison of Precision of all the four techniques used to build the model.

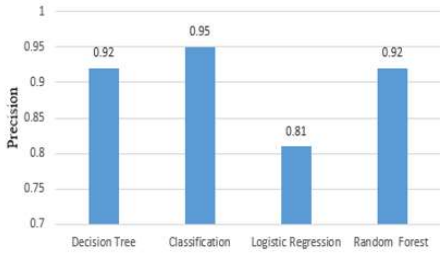

Fig. 5. Comparison of Precision

As we can see from Figure 6 below regarding comparison of Sensitivity of the Techniques. Among all the four techniques applied for generating the model, Classification (KNN) yields the highest Sensitivity (95%). Decision Tree and Random Forest yields the same Sensitivity (92%) while Logistic Regression yields the lowest Sensitivity (82%). Figure below shows the Comparison of Sensitivity of all the four techniques used to build the model.

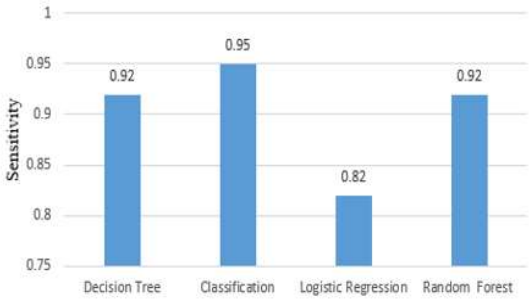

Fig. 6. Comparison of Sensitivity

As we can see from Figure 7 below regarding comparison of RMSE of the Techniques. Among all the four techniques applied for generating the model, Decision Tree and Classification (KNN) yields the lowest RMSE (0.1) followed by Random Forest (2.9). Logistic Regression yields the highest RMSE value (2.9).

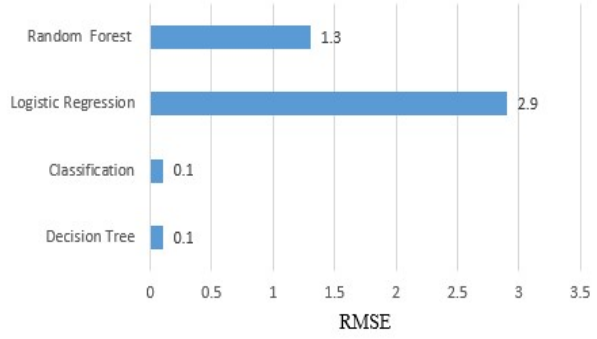

Fig. 7. Comparison of RMSE

In order to assess the performance of our model we considered Accuracy, Sensitivity, Precision and RMSE for evaluating our model. The output revealed that Classification has the highest Accuracy compared to other four techniques. Root mean Square parameter is very important when assessing the performance of the prediction model. It basically reveals the error encountered by prediction model based on the predictor variable. Both Decision Tree and Classification has the lowest RMSE value i.e. (0.1) followed Random Forest (1.3).Logistic Regression yields the highest RMSE value (2.9). Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and is the most important criterion for fit if the main purpose of the model is prediction.

VI. CONCLUSION

Four machine learning techniques were applied to develop the prediction model i.e. Decision Tree, Classification, Logistic Regression and Random Forest. Model evaluation was performed by calculation some of the performance parameters i.e. Accuracy, Sensitivity, Recall and RMSE. Analysis revealed that Classification is best prediction model based on higher accuracy and lower Root Mean Square value. Soil moisture is a deciding factor whether ignition due to Lightening, Debris and Arson will occur or not So far no major research work has been done in order to assess the dryness of soil in the forest land. Even measurement for dryness of soil is possible but acquiring the data to measure the soil dryness over the vast land which covers the densely populated forest is very difficult. Moreover mobilization of resources and budget allocation needs to be taken care of. If data for soil dryness is available then building prediction model based on the dataset of soil dryness would highly contribute for determining the occurrence of Wild fire.

Detail analysis of wind speed direction is needed to further enhance the performance the prediction model which ultimately can predict the category of wild fire and burned area. Other prediction model such as Neural Network and Clustering needs to be applied to predict the wind speed direction, one of the significant and influencing factor for wild fire.

REFERENCES

[1] E. Rashidi, H. Medal, J. Gordon, R Grala, M. Varnerc," A maximal covering location-based model for analyzing the vulnerability of landscapes to wildfires: Assessing the worst-case scenario," European journal of operational research, 258(3), pp.1095-1105, 2017.

[2] Kenward, A., Sanford, T., & Bronzan, J. (2016). WESTERN WILDFIRES.

[3] R. M. Houtman, C. A. Montgomery, A. R. Gagnon, A. R., Calkin, D. E., Dietterich, T. G., McGregor, S., & Crowley, M, "Allowing a wildfire to burn: estimating the effect on future fire suppression costs. International Journal of Wildland Fire, vol. 22(7), pp.871-882, 2013

[4] G. Nyamadzawo, W. Gwenzi, A. Kanda, A Kundhlande, A., & Masona, C., "Understanding the causes, socio-economic and environmental impacts, and management of veld fires in tropical Zimbabwe," Fire science reviews, vol. 2(1), 2013.

[5] R. Wirth, J. Hipp, "CRISP-DM: Towards a standard process model for data mining," In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, pp. 29-39), London, UK: Springer-Verlag, 2000.

[6] S. Xingrong, "Research on time series data mining algorithm based on Bayesian node incremental decision tree," Cluster Computing, vol. 22(4), pp.10361-10370, 2019.

[7] C. Davatzikos, Y. Fan, X. Wu, D. Shen, S. M. Resnick, "Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging," Neurobiology of aging vol. 29(4), pp.514-523, 2008.