Forest fire prediction

Kumuda Benakanahalli Guruprasada Murthy, Raghava Devaraje Urs, and Shiv Kumar Ganesh
Department of Software Engineering, San Jose State University.

Abstract—Forests play an important role in our ecosystem and these vegetation needs to be saved, harnessed for the well-being of all animal life. Forest fire is one of the terrors which can cause massive devastation. These fires may be caused by human beings or by external factors like lightning. Every year millions of dollars are spent by the United States alone to prevent the spread of wildfire. This surely affects the economy of the country battling the wildfires. As the time passes, the continuous occurrence of forest fires causes a drastic change in the climatic conditions. There is an imminent need to control and avoid these outraging fires which destroy the flora and fauna at the same time damage the economy of a country. All the devastation caused by forest fires can be avoided and we can save many lives if the forest fires are detected well in advance so that the authorities can be well prepared to handle them. In this project we mine the weather and wildfire records of the state California, curated from state, federal and local agencies and aim to come up with a model that predicts the occurrence of forest fire. Data mining methodologies like Multiple regression, Support Vector Machine, Random Forest will be used throughout the project to make most accurate forest fire prediction.

Index Terms—Forest fire, Decision trees, Logistic regression, Spacial data, Support Vector Machine, Weather data.

———————————————— ◆ ————————————————

1 INTRODUCTION

FOREST fires are the kind of fires that occur sporadically which eventually become uncontrollable. These fires can be caused with or without human intervention. Though the fires play a vital role in our ecosystem, in 2019, 4.7 million acres [1] of forest was lost due to 50,477 forest fires [1]. The graph 1 below depicts that there has been an increase in the number of forest fires over the past six decades. The fires once instigated, will propagate wildly in all directions causing havoc and threatening the life of the people as well as the wildlife. Though by the looks of it the forest fires are life threatening, a few controlled fires are necessary every now and then to burn up old debris and make way for new healthy vegetation to grow.

Every year countries like the United states of America, Australia and Africa battle forest fires by spending millions of dollars to control them. Thus, a lot of money will be spent on controlling the fire when it could be spent on protecting and improving the vegetation of a country. With this context, it becomes very evident that considering all the factors and actually predicting when and where a forest fire can occur is very beneficial. It not only helps in preserving the flora and fauna but also aids in lifting up the economy of a country. Forest fire prediction can help the authorities to fore see where a fire might occur and take immediate actions like evacuation, depute more firefighters for a particular region, lock down the region, or maybe even start a controlled manual fire called backburning in the same course of the predicted forest fire. This controlled fire will burn up all the dried debris, thereby extinguishing the

fuel (dried fallen leaves, twigs that aid in spreading forest fire) before the uncontrolled forest fire hits. Thus, the forest fire dies out.
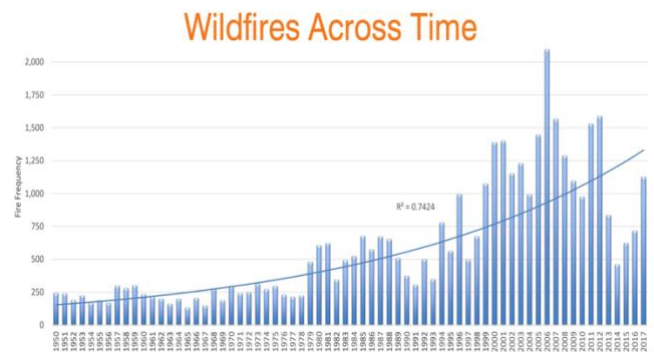

Fig 1: Wildfire history of USA (1950-2017) [2]

2 THE DATA

2.1 Data Collection

The success of a data mining project relies on quality of the data gathered. For this project to predict forest fires, the data set consists of weather and spatial information for different regions within the state of California in the United States. The required weather data was retrieved from *riverside municipal airport, CA-US weather center*. The data collected ranges from April 1998 to current.This station can be found on https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00003171/detail. The spatial wildfire data can be accessed from https://doi.org/10.2737/RDS-2013-0009.4. This is a collection of data procured from the local, state and federal fire organization's resporting systems and it ranges from 1992 to 2020. The data collected contains more than 1.8 million records. This huge data set was filtered to obtain the spatial data for California alone. A python program was

————————————————

- *Kumuda Benakanahalli Guruprasada Murthy is with the San Jose State University, San Jose. E-mail: kumuda.benakanahalliguruprasadamurt@ sjsu.edu.*
- *Raghava Devaraje Urs is with the San Jose State University, San Jose. E-mail: raghavadevaraje.urs@sjsu.edu.*
- *Shiv Kumar Ganesh is with the San Jose State University, San Jose. E-mail: shivkumar.ganesh@sjsu.edu.*

used to query the county name based on geolocation using ArcGIS. The following figure depicts the performed data collection process,
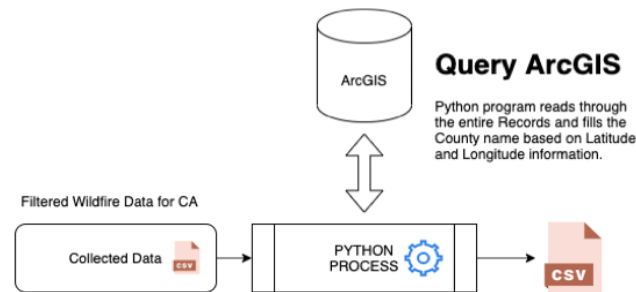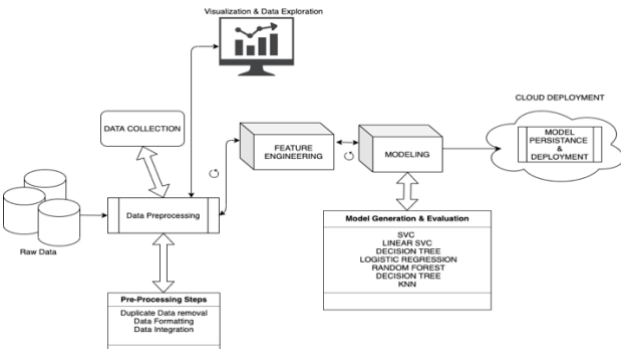

Fig 2: Data collection process

2.2 KDD Architecture

The above figure describes a KDD (Knowledge discovery in database) Architecture which is implemented to perform various data cleaning, processing and model building tasks. This is a typical KDD architecture using which the data is collected, cleaned, preprocessed and then features are engineered. After this process is done, various machine learning algorogitms are used in order to build a suitable classification model which is then deployed on cloud.

2.3 Data cleansing

The data collection process provides us with the final CSV file which has the data related to various counties within California. This gives us the opportunity to narrow down the county with most wildfires. Upon analysing the data, it can be observed that the Riverside county is mostly prone to wildfire in California. The next step would be to filter the data set to differentiate between the data points that represent the days on which forest fire occurred and the days on which there were no fires. Thus, we get a complete data set with both positive and negative data points. Also, this this step we identify the columns from California wildfire data that is not relevant for the analysys.

2.4 Data Pre-Processing

In the pre-processing stage we extract only the relevant information from the given California wildfire data and the weather data. From the previous step with know that the Riverside county has the maximum records of the fire incident, hence we extract the Riverside county data from California wildfire data. Also, sort the data by Date. This enables us to prepare the fire data for merging with

weather data. Next, we add a new column 'IS_WILDFIRE' to the data which helps in classifying a data point as a positive point. From the weather data collected we extract the columns like Maximum temperature, Minimum temperature, Precipiation and Avaerage daily wind speed. We extracted Data such as Active months, time of the day and weekdays as additional features. These columns are more relevant and aid in predicting fires. Also, fill all the negative values for the wildfire occurance with false, extract data till the last available data point in the wildfire data, discard the rest.

2.5 Data normalization

2.5.1 Principal Component Analysis(PCA)

The forest fire data set under consideration has a very large number of features with correlation among few features. PCA aids in dimentionality reduction and helps to remove redundancy by converting high dimentional correlated features into lower dimentional uncorrelated features [3][4]. Following steps are involved in PCA,

[1] Compute mean and covariance for the data set

$$S = \frac{1}{n}\sum_{i=1}^{n}(\vec{x}_i - \vec{x})(\vec{x}_i - \vec{x})^T$$

[2] Singular Vector Decomposition

$$\vec{x} = \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i$$

[3] Target matrix projection

$$S = U\Sigma V^T$$
$$Y = P^T X$$

2.5.2 Feature Scaling

The variables of a data set can be standardized using feature scaling. Min-max scaling is one of the types used of normalization where the values are scaled in range of [0,1] or [-1,1]. After the normalization, the relationship between the original data is preserved[5].

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

x = value to be normalized
min(x) = minimum value that x takes up in a column
max(x) = maximum value that x takes up in a column

2.5.3 Label Encoding

In label encoding every value in the column is represented as a numerical value. We give similar numbers to similar values. The categories with yeas and no are given 1 and 0 respectively. The target variable and its label are encoded along with other features[6].

2.6 Feature engineering

The pre-filtered California wildfire data has 40 different values. Out of which there a multiple IDs that are unique identifiers and increase the dimentionality of the data. Fe columns like OBJECTID, FOD_ID,FPA_ID are purely unique identifier and just increase the dimentionality. Hence, we will remove those values from the data. Also, there are few columns like fire size and fire size class which are very beneficial to corelate the weather data. We can see that when there is no fire, the burn area is zero. Thus, we fill in the burnt area with zero for all the days when there was no fire incident. The weather data did not have the Month and day of the week column, so we ill go ahead and add that. At the end, we fill all the unknown values with the mean of the entire column. At the end of feature engineering, we will have our final data with all the required columns and data points ready to be fed to various model. The correlation matrix(Fig 3) and the pair plot(Fig 4) plotted for the merged dataset shows that there is a high correlation between TMAX(Temperature max), TMIN(Temperature min), TAVG(temperature average) features and precipitation is negatively correlated.
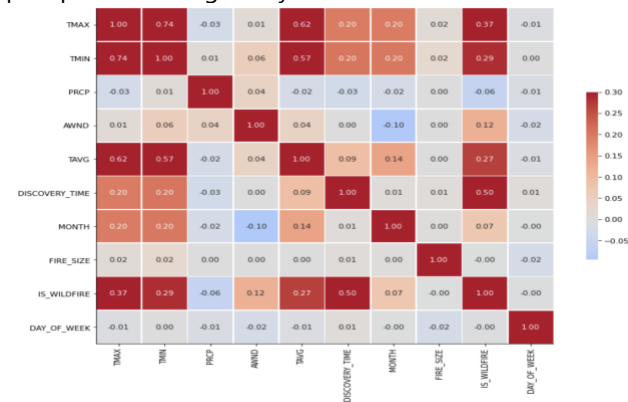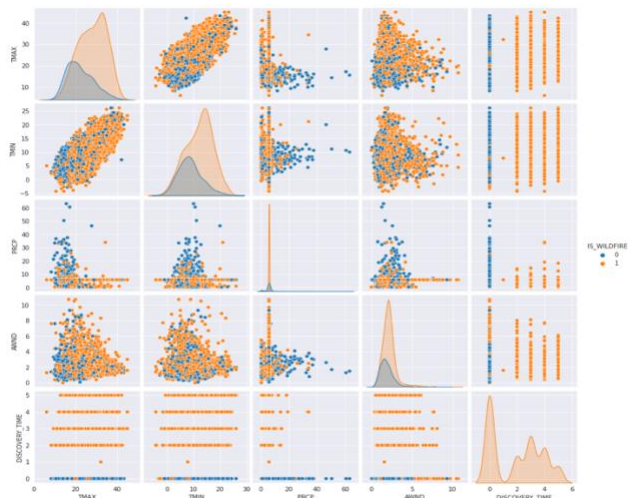

Fig 3: Heatmap correlation matrix

Fig 4: Pair plots

3   MACHINE LEARNING MODELS

3.1 Standardization of the training data

The data obtained from the previous set is then divided in training set and testing set. We consider 67% of the data as the training set and the remaining 33% as test data.

3.2 Model generation

For forest fire prediction we considered the following model generation methods – SVC, LinearSVC, Guassian NB, Decision tree, KNN, Logistic Regression, Random Forest Classifier and SGD Classifier.

3.2.1 Support Vector Machine (SVM)

SVM is one of the highly preferred algorithms as it provides very high accuracy with very low computational power. This method is mainly used for classification task but sometimes used for regression tasks. It transforms
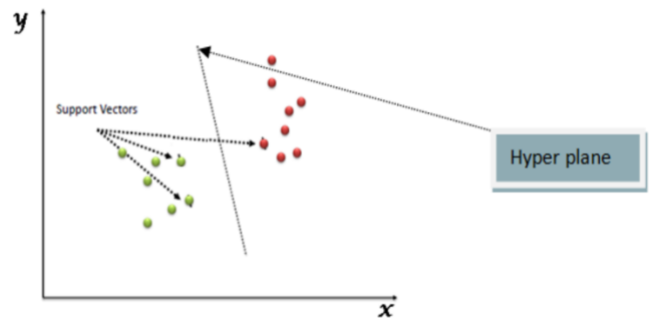

Fig 5: Support vectors and Hyper-plane[8]

training data into higher dimension using nonlinear mapping. A support vector is defined as the distance of the hyperplane from the closest data point. Here, the hyperplane will be in the form of $w^T x + b = 0$[7], where w is weight vector, x is input vector and b is bias. We have used two forms of SVM namely **Support Vector Classifier** and **Linear** **Classifier**.
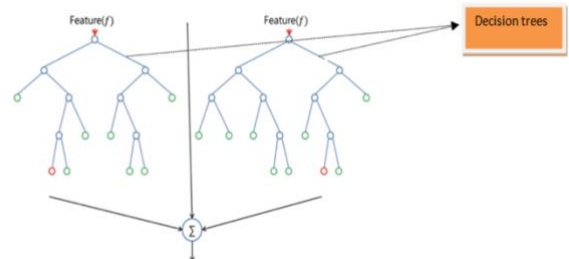
3.2.2 Decision tree

Fig 6: Two decision trees forming a random forest[8]

Decision tree setsup classification models in a tree structure while breaking the data set as small as possible. At the end we will have tree with the nodes representing the decisions and leaf nodes. The following figure shows two decision trees. It takes the output from two trees and gives the final decision.

3.2.3 Logistic regression
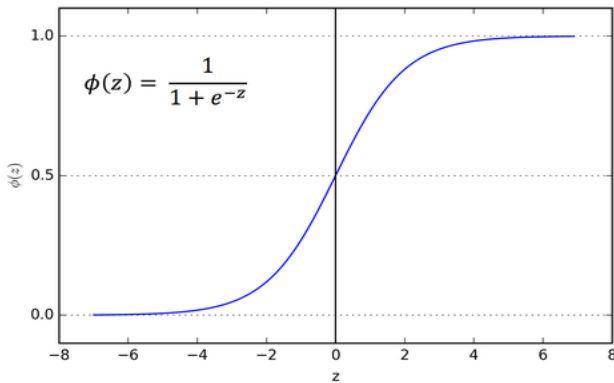

$$\phi(z) = \frac{1}{1 + e^{-z}}$$

Fig 7: Logistic regression sigmoid curve[8]

Logistic regression is the model used for scenarios with binary response value i.e the variables are dichotomous[9]. For prediction usecases like forest fire prediction logistic regression will be easier implement. This model performs well for linearly seperable dataset.

3.2.4 KNN

Logistic regression K-Nearest Neighbour is a simple algorithm and is a part of a vast variety of machine learning algorithm. It's used for both regression and classification problems and it classifies any given data point based on similarity using distance as a measure.

3.2.5 Random Forest

Random forest classifier uses ensemble learning method for classification and regression. It creates multiple decision tree at the time of training. This is a powerful learning method.

3.3 Model evaluation and results

For forest fire prediction, all the above explained machine learning models were applied and the results are represented in the following tables.

Model type	Accuracy
SVC	0.83
Linear SVC	0.82
Decision Trees	0.77
KNN	0.76
Logistic Regression	0.82
Random Forest	0.82

Table 1: Models and their accuracy

| Model type	Precision	Recall	F1 score	Support
SVC	0.79	0.80	0.79	1979
Linear SVC	0.77	0.78	0.78	1979
Decision Trees	0.71	0.72	0.72	1979
KNN	0.73	0.78	0.74	1979
Logistic Regression	0.77	0.78	0.78	1979
Random Forest	0.78	0.79	0.78	1979

Table 2: Models and their classification report

From the above tables it is clear that Support Vector Machine performs best and gives the best accuracy for forest fire prediction.

3.4 Data visualisation

Exploratory Data Analysis was being done on the entire wildfire dataset. Some inferences derieved were really important ones. We can see that these inferences show a heavy correlation towards the month and the cause of the wildfire.

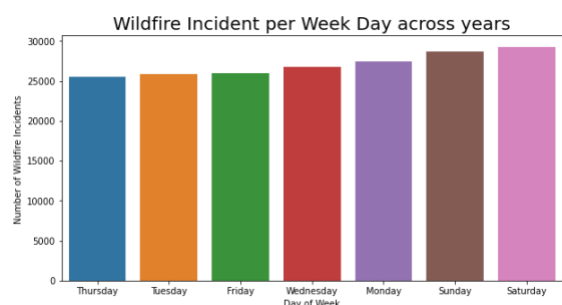### 3.5.1 Causes of wildfire
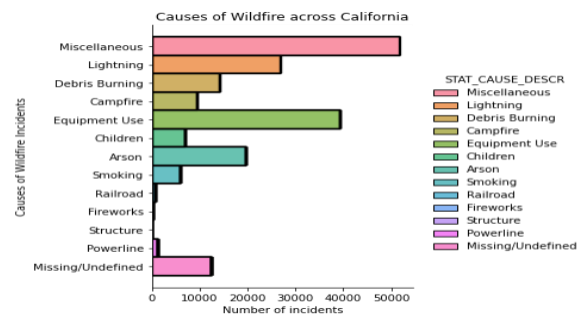
We can see that the the major cause of wildfire after miscellaneous causes was equipment usage and lightening.

3.5.2 Most active months
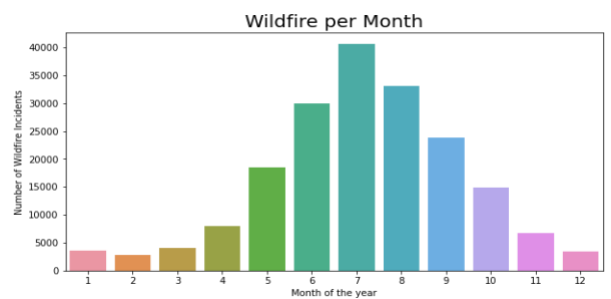
We see that the wildfire incidents drastically increase from the month of May to July and then start reducing. These are typically the summer months when people are out in nature.

3.5.3 Most active days

One can also observe that the number of incidents gradually increase during the weekend thereby symboling a huge involvement of human factor.

4 CONCLUSION

From our analysis and model comparison we see that SVC provided us with a better accuracy and F1 score. This was indicative that the algorithm was performing better than the others. Models generated from the above activity can be used for prediction of wildfire in Riverside County whereas the same methodologies and pipeline can be setup at other locations in order to predict the likelihood of a wildfire event.

ACKNOWLEDGMENT

REFERENCES

[1] Congressional Research Service, September 1, 2020 [https://fas.org/sgp/crs/misc/IF10244.pdf]

[2] https://giscenter.isu.edu/research/Techpg/nasa_RECOVER/pdf/Geogr aphyWildfires.pdf

[3] Chandra Paul, Liton & Suman, Abdulla & Sultan, Nahid, "Methodological analysis of principal component analysis (PCA) method", International Journal of Computational Engineering & Management, 16. 32-38, 2016

[4] Jonathon Shlens," A Tutorial on Principal Component Analysis", Educational, 51, 2014

[5] S.Gopal Krishna Patro & Kishore Kumar Sahu, "Normalization: A Preprocessing Stage", IARJSET, 10.17148/IARJSET.2015.2305, 2015

[6] Kedar Potdar, Taher Pardawala, & Chinmay Pai, "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers", International Journal of Computer Applications. 175. 7-9. 10.5120/ijca2017915495, 2017

[7] Piyush Jain, Sean C.P. Coogan, Sriram Ganapathi Subramanian, Mark Crowley, Steve Taylor, and Mike D. Flannigan. A review of machine learning applications in wildfire science and management. Environmental Reviews. 28(4): 478-505. https://doi.org/10.1139/er-2020-0019

[8] R. Rishickesh, A. Shahina, A. Nayeemulla Khan, "Predicting Forest Fires using Supervised and Ensemble Machine Learning Algorithms", IJRTE, DOI: 10.35940/ijrte.B2878.078219, July 2019.

[9] Joanne Peng, Kuk Lida Lee, & Gary M. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting", Journal of Educational Research - J EDUC RES. 96, 3-14. 10.1080/00220670209598786, 2002

[10] G. E. Sakr, I. H. Elhajj, G. Mitri and U. C. Wejinya, "Artificial intelligence for forest fire prediction," 2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, Montreal, ON, 2010, pp. 1311-1316, doi: 10.1109/AIM.2010.5695809.

[11] N. Hamadeh, A. Hilal, B. Daya and P. Chauvet, "Studying the factors affecting the risk of forest fire occurrence and applying neural networks for prediction," 2015 SAI Intelligent Systems Conference (IntelliSys), London, 2015, pp. 522-526, doi: 10.1109/IntelliSys.2015.7361189.