Forest Fire Prediction using Data Science

Aafia Iqbal
2018A7PS0061U

Objective

1. Predict the main month of Forest Fires in the northeast regions of Portugal.
2. Apply and evaluate what is the best algorithm for predicting Forest Fires

Introduction

Data Science is a field that studies Structured or Unstructured data and extracts information from the data and analyzes and finds valuable information. Data Science algorithms look for patterns and predict the outcomes. Data Science can be very helpful in avoiding and taking precautions for the environment.
Forest Fires have caused major damages to the environment. Data Science can be applied in predicting the main month these Forest Fires depending on particular factors of that region. There can be one reason or many reasons for the cause of Forest Fires, like the weather of the region, temperature, closeness to inhabitants or industrial land, month, temperature etc. These factors can be analyzed and evaluated region by region to prevent large damage to the Environment and Wildlife.
It was reported that the Forest Fires of 2017 in Portugal burnt nearly 500,000 hectares.[11] This fire highly affected the environment and economy and brought great loss to human lives. These forest fires also led to drastic increase in the temperatures in Portugal. Portugal faced its first wave of Forest Fires 2017 in June , these fires of June have been reported to be the main cause of the bigger Forest Fires which happened in October in the same year, 2017. The main cause of these Forest Fires of 2017 in Portugal was assumed to be the climatic conditions.
In our Project we tend to analyze and study the season of Forest Fires of the northeastern regions of Portugal. We apply various Data Science Algorithms in predicting the Forest Fires of northeastern regions of Portugal. Then we apply evaluation metrics to analyze which algorithm gives the most satisfactory results.

Data Set

https://archive.ics.uci.edu/ml/datasets/forest+fires

X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0	0
7	4	oct	tue	90.6	35.4	669.1	6.7	18	33	0.9	0	0
7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0	0
8	6	mar	fri	91.7	33.3	77.5	9	8.3	97	4	0.2	0
8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0	0
8	6	aug	sun	92.3	85.3	488	14.7	22.2	29	5.4	0	0
8	6	aug	mon	92.3	88.9	495.6	8.5	24.1	27	3.1	0	0
8	6	aug	mon	91.5	145.4	608.2	10.7	8	86	2.2	0	0
8	6	sep	tue	91	129.5	692.6	7	13.1	63	5.4	0	0
7	5	sep	sat	92.5	88	698.6	7.1	22.8	40	4	0	0
7	5	sep	sat	92.5	88	698.6	7.1	17.8	51	7.2	0	0
7	5	sep	sat	92.8	73.2	713	22.6	19.3	38	4	0	0
6	5	aug	fri	63.5	70.8	665.3	0.8	17	72	6.7	0	0
6	5	sep	mon	90.9	126.5	686.5	7	21.3	42	2.2	0	0
6	5	sep	wed	92.9	133.3	699.6	9.2	26.4	21	4.5	0	0
6	5	sep	fri	93.3	141.2	713.9	13.9	22.9	44	5.4	0	0
5	5	mar	sat	91.7	35.8	80.8	7.8	15.1	27	5.4	0	0
8	5	oct	mon	84.9	32.8	664.2	3	16.7	47	4.9	0	0
6	4	mar	wed	89.2	27.9	70.8	6.3	15.9	35	4	0	0

Literature Survey

Guoli Zhang et al discuss the use of Conventional Neural Networks for prediction of Forest Fire Susceptibility in Yunnan Province, China. They have taken data between the years 2002-2010 for considerations and also 14 factors that influence Forest Fires. They then applied the data and practised it on various statistical methods like Wilcoxon signed-rank test, the ROC curve, and area under the curve . They used various different algorithms like Random forest, SVM, multilayer perceptron neural network, and kernel logistic regression classifiers. And finally came to a conclusion that the CNN Model gives the most accurate results.[1]

In [2], Volkan Sevinc et al. discuss the application of Bayesian network Model for predicting the causes of Forest Fires. They collected data between 2008-2018 of the Mugla Regional Directorate of Forestry, Turkey. They took various different factors into consideration like, temperature, velocity of wind, month, distance from inhabitants, amount of burnt area, distance from man-cultivated land, distance from roads and tree species. They finally came to the conclusion that the month is the most important factor for causing Forest Fires.

Mahyat Shafapour Tehrany et al. discuss the use of LogitBoost ensemble-based decision tree in predicting Forest Fires. They collected data of Lao Cai Region, Vietnam. They applied the data on various algorithms like Area under the curve , the Kappa index, accuracy, specificity, (PPV), sensitivity, and (NPV). They finally came to the conclusion that LEDT (LogitBoost ensemble-based decision tree) provides the best and accurate results. [3]

In [4], Binh Thai Pham et al discuss which algorithm is the best for determining and predicting Fire susceptibility. They collected data from Pu Mat National Park, Vietnam. The data consisted of 57 previous forest fires and various reasons for the forest fire like level of elevation, degree of slope, average annual temperature, index of drought, river density, area of land, and distances

from roads and inhabitants. They applied the data on 4 algorithms Bayes Classifier , Naïve Bayes Classifier, Decision Tree , and Multivariate Logistic Regression. Among these 4 algorithms Bayes Classifier showed the most appropriate results on Area Under the Curve.

Haoyuan Hong et al created a Forest Fire susceptibility mapping system using Data Mining techniques like Random Forest Classifier and Support Vector Machine. These techniques were evaluated using the ROC curve. The authors collected data from Dayu County, China. This data consisted of factors like level of elevation, angle of slope, land use(commercial or residential), soil cover, heat load index, vegetation index, average annual temperature, average annual wind speed, average annual rainfall, distance to the nearest water body and distance to inhabitants. It was noticed that the Random Forest Classifier gave the best results. [5]

In [6], Haifeng Lin et al discuss Fuzzy Logic Algorithms and Big Data Algorithms for Forest Fire detection. They collected data from Nanjing City region, China. They concluded that Human Behavior and the Weather were the most important factors for Forest Fires.

Ljubomir Gigović et al use various Data Science techniques like Random Forest Classifiers and Support Vector Machines. They evaluated these techniques using the ROC-AUC curve. The study was done on the basis of data collected from Tara National Park, Serbia. Though both the techniques showed nearly very close values on the ROC-AUC curve but the SVM method was more accurate. [7]

In [8], Dieu Tien Bui et al introduce a new Machine learning approach for understanding the causes of Forest Fires. They collected data from the Lao Cai province, Vietnam. The authors established a GIS database and collected data consisting of 10 factors for predicting Forest Fires. They proposed the Multivariate Adaptive Regression Splines - Differential Flower Pollination Method. The authors concluded that their proposed method showed higher performance of the ROC-AUC curve and few more evaluation methods as compared to other Machine Learning Techniques.

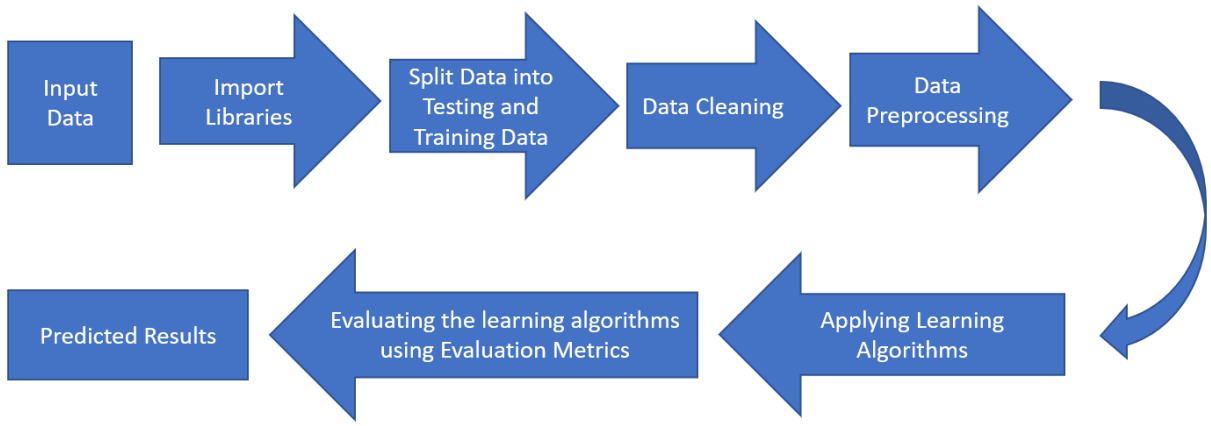
Hung Van Le et al propose a new Machine Learning Technique namely the Differential Flower Pollination - mini-match back propagation method for predicting Forest Fires. The authors worked on data collected from Lam Dong province, Vietnam. The proposed technique showed satisfactory results as compared to other Machine Learning techniques. [9]

In [10], Hamid Reza Pourghasemi et al discuss Forest Fire Susceptibility. They collected data from FARS Province, Iran. The authors used algorithms like Mixture Discriminant Analysis, Boosted Regression Tree and General Linear Model. They evaluated these ALgorithms using the ROC curve. The Boosted Regression Tree method showed the most satisfactory results among them.

Methodology

This section includes the proposed methodology for Forest Fire Prediction using Data Science.

The dataset includes information like month, day, FFMC index, DMC index, DC index, ISI index, temperature, relative humidity, wind speed, level of rain, and the burnt area. All these factors are used for predicting Forest Fires.
We apply various Data Science Algorithms like Naive Bayes Classifier, K-Means, Random Forest Classifier, and Logic Regression to our dataset , then we evaluate the results using various evaluation metrics like Accuracy , Precision , Specificity and Sensitivity.



Block Diagram explaining methodology

This section is further divided into 2 sections , Section 1 explains the algorithms being applied, Section 2 comprises the evaluation metrics required for predicting Forest Fires.

Section 1 Algorithms

This section is further subdivided into 4 parts , each describing the algorithms being applied.

Naive Bayes Classifier

Naive Bayesian Classification is broadly based on the Bayes Theorem. The Bayes Theorem acts when an event has already occurred thus giving the conditional probability of another event. Naive Bayesian Classification is a Supervised Machine Learning Technique.

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

Equation for Naive Bayes Classifier

K-Means Clustering Algorithm

K means is a partitional clustering algorithm. The data is initially clustered into groups based on similarity. The “K” refers to the number of Clusters. Initially random distinct points are chosen as the Centroid. Then further more points are chosen and clustered with the Centroids depending on the level of similarity with the Centroid. K- Means Clustering Algorithm is classified as Unsupervised Machine Learning Technique.

Random Forest Classifier

Random Forest Classification is an Ensemble Learning Classifier. It works on the basis of Decision Trees created in the training Phase. Random Forest Classification is a Supervised Machine Learning Technique

Logic Regression

Logic Regression works on Continuous data and gives results in binary form, i.e, True or False. It is classified as Supervised Machine Learning Technique.

$$\log\left(\frac{Y}{1-Y}\right) = C+B1X1+B2X2\ldots$$

Equation for Logistic Regression

Section 2 Evaluation Metrics

Accuracy

Accuracy is the ratio of sum of Correctly predicted Positive and correctly predicted negative results to the sum of the total results taken in consideration.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{False Positives} + \text{False Negatives} + \text{True Negatives}}$$

Equation for Accuracy

Precision

Precision is the ratio of correctly predicted positive results to the sum of correctly predicted positive and wrongly predicted positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Equation for Precision

Experimental Setup

The Forest Fires Dataset contains 517 entries. We use Jupyter Notebook 6.1.4 using Anaconda open source network in the language Python 3.8.
We import Python Libraries like Pandas, Sklearn, Numpy, matplotlib.pyplot, and pickle.

The data considered for the experiment contains 13 attributes as follows: X-axis spatial coordinate, Y-axis spatial coordinate, Month, Day, FFMC, DMC, DC, ISI, Temperature, RH, Wind, Rain, and Area.

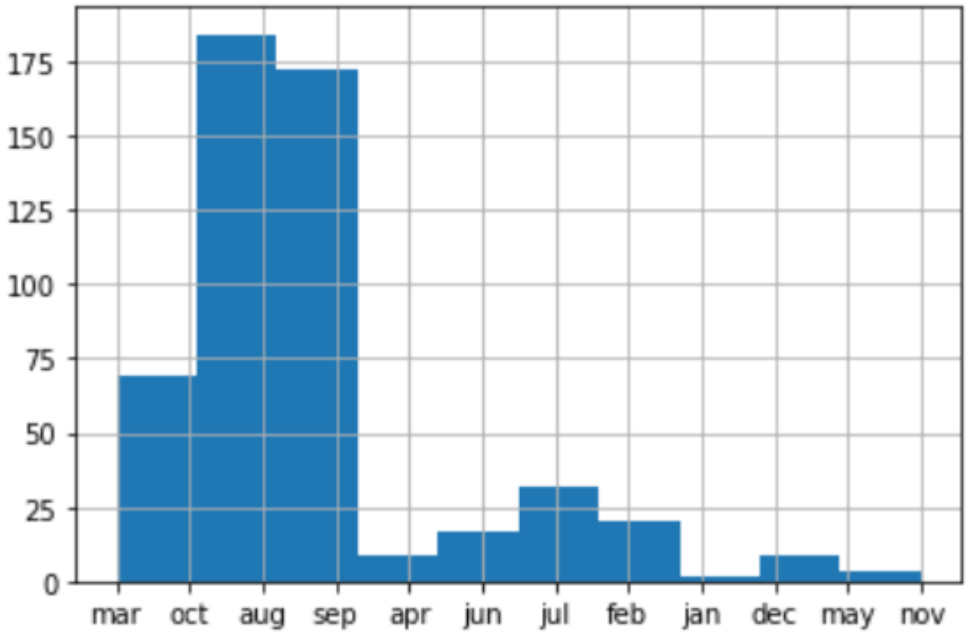
We apply four algorithms - Naive Bayes Classifier, K-Means, Random Forest Classifier, and Logic Regression on our dataset forestfires.csv.

Each algorithm is trained and tested to give appropriate values of Predicted and Actual Data. Further Accuracy and Precision of each algorithm is recorded. Finally the accuracy and precision of all the algorithms is compared to find the best algorithm for predicting Forest Fires.

Results and Discussion

In this section we discuss the experimental results we obtained after executing our code
In Section 1, we discuss the occurrence of Forest Fires. In Section 2, we discuss the results obtained after applying the four algorithms and evaluating them on the basis of few evaluation metrics.

Section 1
There are 517 entries of Data. The data comprises 12 attributes. Our first objective is to find the month in which forest fires are observed in a high number.



Graph 1. Histogram depicting the months in which the forest fire was observed.

In the above graph, we notice that most of the Forest Fires occurred between the months of August and October. This is the end of the Summer season in Portugal.
We also notice that the least number of Forest Fires can be noticed between the months January to February, this is Winter season in Portugal.

Section 2

Logistic Regression

We first apply Logistic Regression. We import the model
`from sklearn.linear_model import LogisticRegression.`

Then we train the data on the Logistic Regression Classifier. We also test the data to obtain how many Forest Fires were predicted and how many actually occurred.
We apply evaluation metrics like Accuracy and Precision on the obtained data.

Accuracy	0.4807692307692308
Precision	0.5079365079365079

Table . Results of Logistic Regression

Random Forest Classifier

We import the package `from sklearn.ensemble import RandomForestClassifier.`

Then we feed that data to the classifier and train it. We also test the data to obtain how many Forest Fires were predicted and how many actually occurred.

We apply evaluation metrics like Accuracy and Precision on the obtained data.

Accuracy	0.5576923076923077
Precision	0.5737704918032787

Table . Results of Random Forest Classifier

Naive Bayes Classifier

We import the package `from sklearn.naive_bayes import GaussianNB.`

Then we feed that data to the classifier and train it. We also test the data to obtain how many Forest Fires were predicted and how many actually occurred.

We apply evaluation metrics like Accuracy and Precision on the obtained data.

Accuracy	0.46153846153846156
Precision	0.3333333333333333

Table . Results of Naive Bayes Classifier

K Means

We import the package `from sklearn.cluster import KMeans`

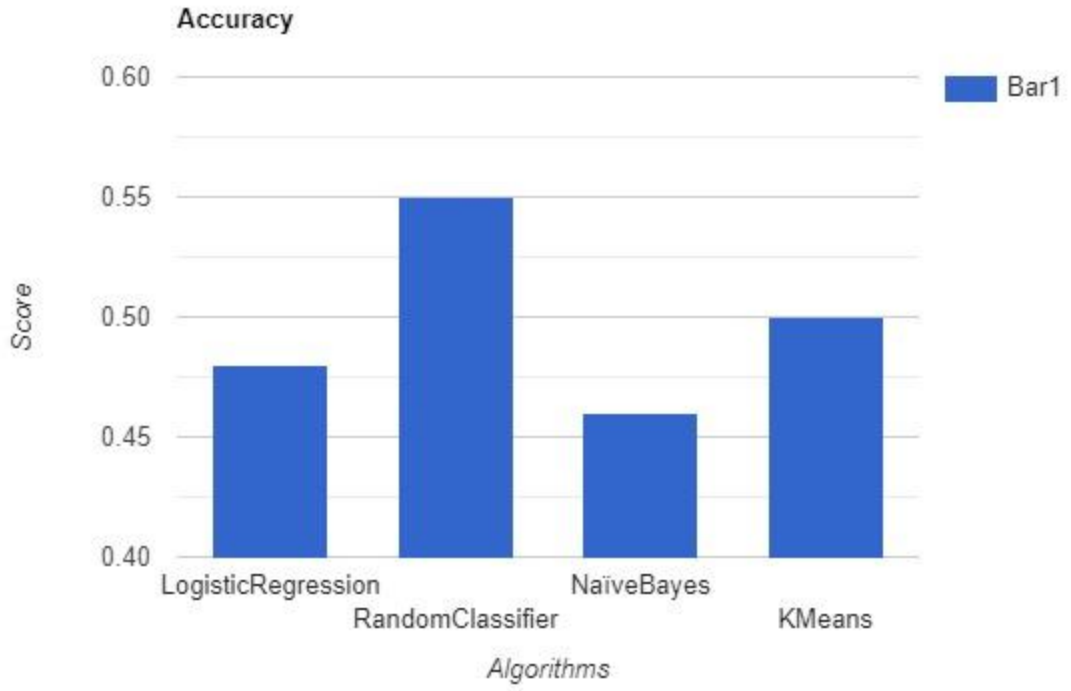
Then we feed that data to the classifier and train it. The data is processed on the basis of 4 factors Rain, Temperature, RH and Wind.

After training and testing the data, we evaluate it using Accuracy and Precision evaluation Metrics. Finally we plot the centroids of the clusters and obtain a graph.

Accuracy	0.5
Precision	0.5483870967741935

Table . Results of Naive Bayes Classifier

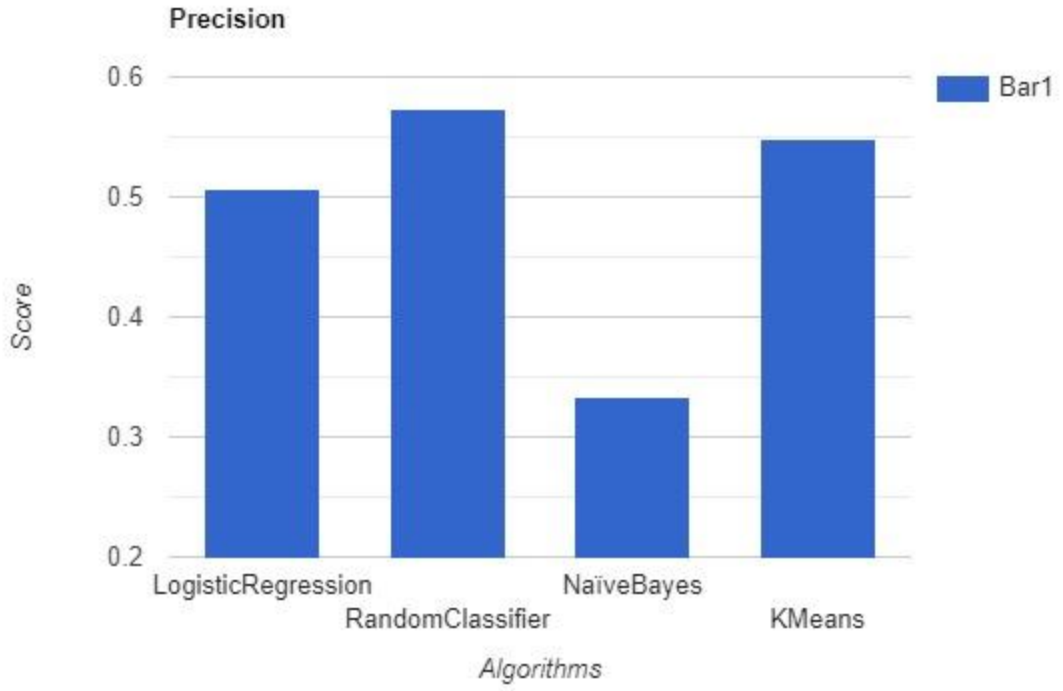
Accuracy



Graph 2 . Comparison of Accuracy of algorithms applied

We notice that Random Forest Classifier gives the most accurate results. The accuracy of the algorithms in descending order: Random Forest Classifier, K-means, Logistic Regression and Naive Bayes Classifier.

Precision



Graph 3. Comparison of Precision of algorithms applied

We notice that Random Forest Classifier gives the most precise results. The precision of the algorithms in descending order: Random Forest Classifier, K-means, Logistic Regression and Naive Bayes Classifier.

Conclusion

Forest Fires have led to major damages environmentally as well as industrially. Many people and animals have lost their homes to forest fires. Timely interventions can help prevent these forest fires or at least reduce the impact of the damage caused.

We studied and analyzed the dataset and found out that most of the forest fires occurred during the time period between August to October. The government of Portugal can be precautious and be extremely alert in this time period.

We applied four different Algorithms - Naive Bayes Classifier, K-Means, Random Forest Classifier, and Logic Regression to our dataset. We found out that Random Forest Classification is the best algorithm in predicting Forest Fires. Random Forest Classifier gives the most accurate and precise results as compared to the other algorithms.

References

1. Zhang, Guoli, Ming Wang, and Kai Liu. "Forest fire susceptibility modeling using a convolutional neural network for Yunnan province of China." *International Journal of Disaster Risk Science* 10.3 (2019): 386-403.
2. Sevinc, Volkan, Omer Kucuk, and Merih Goltas. "A Bayesian network model for prediction and analysis of possible forest fire causes." *Forest Ecology and Management* 457 (2020): 117723.
3. Tehrany, Mahyat Shafapour, et al. "A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using logitboost machine learning classifier and multi-source geospatial data." *Theoretical and Applied Climatology* 137.1 (2019): 637-653.
4. Pham, Binh Thai, et al. "Performance evaluation of machine learning methods for forest fire modeling and prediction." *Symmetry* 12.6 (2020): 1022.
5. Hong, Haoyuan, et al. "Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models. The case of Dayu County, China." *Science of the total environment* 630 (2018): 1044-1056.
6. Lin, Haifeng, et al. "A fuzzy inference and big data analysis algorithm for the prediction of forest fire based on rechargeable wireless sensor networks." *Sustainable Computing: Informatics and Systems* 18 (2018): 101-111.
7. Gigović, Ljubomir, et al. "Testing a new ensemble model based on SVM and random forest in forest fire susceptibility assessment and its mapping in Serbia's Tara National Park." *Forests* 10.5 (2019): 408.
8. Bui, Dieu Tien, Nhat-Duc Hoang, and Pijush Samui. "Spatial pattern analysis and prediction of forest fire using new machine learning approach of Multivariate Adaptive Regression Splines and Differential Flower Pollination optimization: A case study at Lao Cai province (Viet Nam)." *Journal of environmental management* 237 (2019): 476-487.
9. Bui, Dieu Tien, Hung Van Le, and Nhat-Duc Hoang. "GIS-based spatial prediction of tropical forest fire danger using a new hybrid machine learning method." *Ecological Informatics* 48 (2018): 104-116.
10. Pourghasemi, Hamid Reza, et al. "Application of learning vector quantization and different machine learning techniques to assessing forest fire influence factors and spatial modelling." Environmental research 184 (2020): 109321.
11. Turco, Marco, et al. "Climate drivers of the 2017 devastating fires in Portugal." *Scientific reports* 9.1 (2019): 1-8.