Predicting forest fires: Advanced Data Analysis 2021

Dimosthenis PAPADIMITRIOU
HEC Lausanne
dimosthenis.papadimitriou@unil.ch

Boris VON SIEBENTHAL
HEC Lausanne
boris.vonsiebenthal@unil.ch

***Abstract*—This project aims at using an existing historical dataset of past forest fires in order to train a model to try and predict future occurrences of wild forest fires. By using different supervised learning methods, we analyse the efficiency of our models and compare them to each other to identify which models work best to accurately figure out whether or not there will be a forest fire.**

I. Introduction

Following last world war, Europe's forests and grasslands expanded by approximately 150,000 square kilometers. Forest fires (a word used in Europe to describe unintentional fires that burn forests and wildlands) are a serious issue. Frequently misunderstood as a problem affecting virtually only Mediterranean countries, this paper Forest Fires in Europe: Facts and Challenges show specifically how fire is now a hazard that affects most of the European countries. Especially, in Portugal, we have a Chilean pattern of geographical repetition, a dramatic scenario that unfolds when cold Atlantic winds meet a hot Mediterranean summer. Although, Chile does not belong in Mediterranean geographically, it belongs in the same climate zone (Sub-tropical), which provides it with a lot of common factors

A large portion of land has been handed over to eucalyptus mono-cultures in the sparsely inhabited north Portugal and frequently inaccessible interior areas, and vast abandoned woods may be seen over the nearby mountains. It is easy to observe that the majority of what is burned is eucalyptus plantations and unchecked pine forests, a multi-faceted problem that must be addressed. This small geographic part of Portugal does remind us that wildfires pose a serious problem regarding that four initial deadly wildfires that erupted across northern Portugal in the afternoon of 17 June 2017 resulted in at least 66 deaths and 204 injured people.

Forests have an important ecological impact because they filter water through soil columns as they pose as infiltrators (Bredemeier et al., 2011). It is also highly known that forests consist of a natural obstacle in possible flooding (Farley et al., 2005) (Schüler et al., 2006). There is no question that actions should be developed to help design fire prevention strategies by implementing statistical data with infrastructure investments, such as soil erosion and sediment transfer to downstream environments, as concluded in (Grangeon Thomas et al., 2021). The best collection of references, data, knowledge, and comparison between fire events was found in the "Wildfire Hazards, Risks, and Disasters", a serious collection of papers that analyze different economic, social, geographical, and cultural aspects and damages, each one based on a different target group. We are going to take some of the examples found in there to underline the problem.

Fig. 1. NASA satelite view of the 2017 Northern Portugal wildfires

The United States maintains a large database of forest fire records that is used to renew applications for combating wildfires, particularly in California. The expanding amount of acres burnt each year, as well as the expanding number of people living in or near fire-prone areas, make wildfire management a more complicated and difficult challenge to solve. Given the importance of social concerns in shaping present challenges and defining future routes, a thorough grasp of social dynamics will be essential. In Europe, on the other hand, the causes of forest fires are yet unknown. (Leone et al.,2009)(Long et al., 2009)(Lovreglio et al., 2006). In 1933, a survey of fires produced by the Institut Internatinal d'Agriculture (I.I.A., 1933) showed that fire causes were mainly related to the negligent use of fire in agriculture, railroad sparks, coal kilning in forests, powerlines, and to a much lesser extent, voluntary actions.

Furthermore, the original paper (Cortez and Morais, 2007), was of great work and contains data of forest fires in a

highly affected area in Northern Portugal, the Montesinho natural park. As the paper claims several Data Mining methods were applied. The output 'area' was first transformed with an ln(x+1) function. After fitting the models, the outputs were post-processed with the inverse of the ln(x+1) transform. Four different input setups were used. The experiments were conducted using 10-fold (cross-validation) x 30 runs. Two regression metrics were measured: MAD and RMSE. A Gaussian support vector machine (SVM) fed with only 4 direct weather conditions (temp, RH, wind, and rain) obtained the best MAD value: 12.71 +- 0.01 (mean and confidence interval within 95% using at-student distribution). The best RMSE was attained by the naive mean predictor. An analysis of the regression error curve (REC) shows that the SVM model predicts more examples within a lower admitted error. In effect, the SVM model predicts better small fires, which are the majority.

Also, the given paper that provides the dataset uses automatic tools based on local sensors provided by meteorological stations. These meteorological metrics like temperature and wind are known to influence forest fires and several fire indexes, such as the Forest Fire Weather Index (FWI). Based on the abstract, they try to assimilate data mining processes to predict the burned area of forest fires. They use five DM techniques like Support Vector Machines (SVM) and Random Forests, and they test four distinct feature selection setups (using spatial, temporal, FWI components, and weather attributes) on recent real-world data collected from the northeast region of Portugal. The best configuration uses an SVM and four meteorological inputs (temperature, relative humidity, rain, and wind) and it is capable of predicting the burned area of small fires, which are more frequent. They conclude that this is particularly useful for improving firefighting resource management. Despite the high complexity in the original paper, we will try to use our knowledge and some known applications in Machine Learning to re-create a model that will tell us how close it is to predict or evaluate those fires from a statistical approach and if these data have a true meaning to the fire events.

II. Description of the research question and the relevant literature

A. Description of the research question

This paper is intended to use an existed public dataset from forest fires to try to predict possible fire events or find the correlation between different variables of those events.

B. Relevant literature

III. Methodology

To approach this question, we assessed that the crucial part would be to find a suitable dataset to base our model upon. This essentially represented an important step that would majorly impact the results we would later be able to gather. We first looked on a broad scale looking for a suitable database in the whole of Europe, we believed that the benefits of having the European Union as a starting point would grant us easier access to gather a centralized database that would contain more entries than any other individual country built dataset. Unfortunately, this proved more complicated than we initially anticipated. As a matter of fact, we struggled to find complete, up to date, data; the main barrier we faced, was the lack of an accessible dataset to the public, from what we could tell, the data were indeed collected and were somewhat up to date, nevertheless, their public access was either limited or only accessible via an already processed form such as heatmaps, graphs and other visual interpretation. After various attempts at gathering our dataset, we eventually managed to find what we were looking for. In order to obtain the data, we needed, we were constrained to look into older data collection, which seemed easier to access than more recent datasets. We settled for a local dataset from a natural park located in Portugal. Our

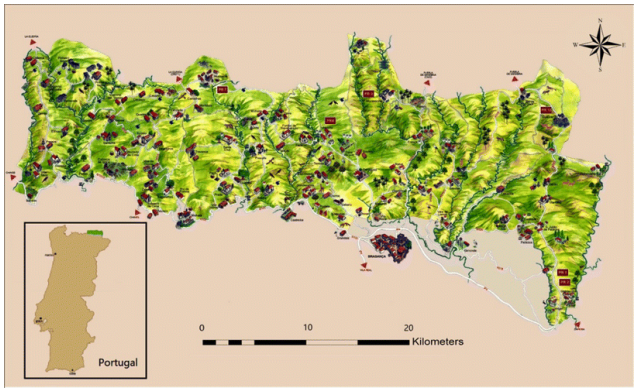

Fig. 2. Map of the natural Park of Montesinho

collection of forest fire data are from the Montesinho natural park, from the northeast region of Portugal. This park contains a high flora and fauna diversity. Portugal is mostly embedded into the upper Mediterranean climate, the average annual temperature is within the range of 8 to 12◦C. The data used in the experiments was collected over the time period starting in January 2000 and spanning until December 2003. The database consists of a collection of fire occurrences gathered by the inspector that was responsible for the Montesinho region. On a daily basis, every time a forest fire occurred, several features would be registered, such as the time, date, spatial location, the type of vegetation involved, the six components of the FWI system and the total burned area. These data were merged with the database was collected by the Braganc, a Polytechnic Institute, containing several weather observations that were recorded by a meteorological station located in the centre of the Montesinho park. The database consists of 517 entries.

A. Forest Fire data clarifications

The forest Fire Weather Index (FWI) is the Canadian system for rating fire danger and it includes six components: Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI) and FWI. The first three are related to fuel codes: the FFMC denotes the moisture content surface litter and influences

ignition and fire spread, while the DMC and DC represent the moisture content of shallow and deep organic layers, which affect fire intensity. The ISI is a score that correlates with fire velocity spread, while BUI represents the amount of available fuel. The FWI index is an indicator of fire intensity and it combines the two previous components. Although different scales are used for each of the FWI elements, high values suggest more severe burning conditions.

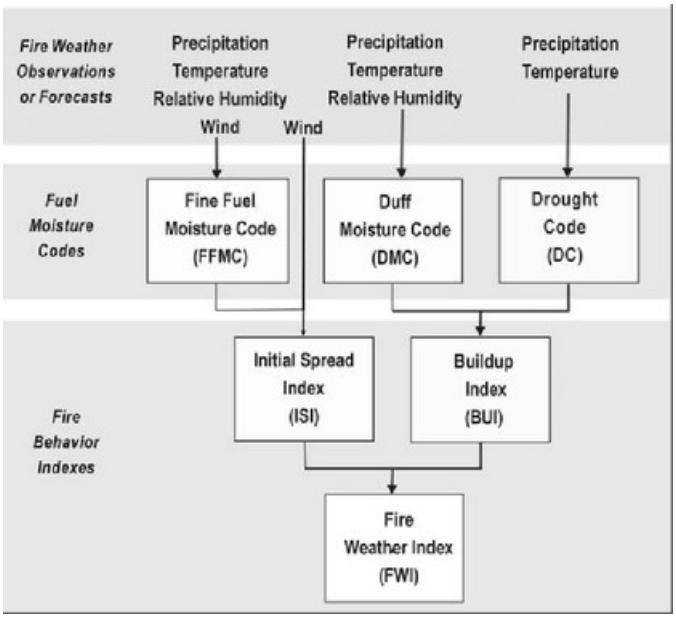

Fig. 3. FWI system

With this dataset, we would settle within this time frame and try to figure out whether we could gather enough data to attempt at building and training a classifier that would, in turn, be able to identify and recognize patterns in the training sample and then be able to predict accurately enough the future occurrences of wildfires. We decided to select different classifier models to compare them together and observe which one would yield better results.

B. Description of the regressors selected

These are the regression models that we selected:

- **Ridge Regression:** Ridge regressions are commonly used when we have to deal with multicollinearity, this is something that we have to face here as the different predictors are correlated with one another, we will try to fit the Ridge Regressor in order to see whether it is able to return an accurate prediction of the burnt area.
- **Gaussian Process Regression:** We also implement a gaussian process regression model, they usually perform better than others on a smaller dataset, since the Bayesian approach attempts to infer the probability distribution of the sample instead of trying to reach exact values for all parameters, so it might help reach a better accuracy.
- **Support Vector Machine Regression:** The strength of the SVM regression model is that we can do a regression to find non-linear relations on the predicted variable thanks to the Radial-basis function kernel (RBF).
- **Random Forest Regression:** A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees. The strength of the Random forest regression is that it fits many problems and provide flexibility regarding non-linear relationships between variables. We expect it to run better than the other models.

C. Description of the classifier selected

Now let's talk about the classification models that we used in order to obtain predictions on the actual occurrence of future fires:

- **Support Vector Machine Classification:** The SVM classifier works better when data are hardly separable via linear classification, since we are not sure about the inferences on our database, the SVM is a good classifier to attempt to identify patterns in the different input features.
- **Random Forest Classification:** The Random forest algorithm is a powerful one as a high number of uncorrelated individual predictions based on decision trees are made and the result that the model accepts is one that the majority of the individual trees predicted.

D. Code structure

The procedure is the following :

1) Split the data samples into training data and test data
2) Train the classifier on the training data sample
3) Test the classifier on the test data split
4) Compute the scoring and error measures to assess the efficiency
5) Compare all the models together and review the results

IV. Implementation of the code

Before we start on the code, we should define the libraries that were used in the process of making this code:

A. Notable Libraries

- **Sklearn:** Scikit-learn is an open-source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection and evaluation, and many other utilities. It possess a lot other sub-libraries categorized in relation to the different tools needed at each steps of the construction of a supervised learning model. This library is the foundation of the supervised learning models that we constructed for this project. It provided us with all the necessary methods to be able to create the model, preprocess the sample data, train the different models and showcase the results.
- **Pandas:** The Pandas library is intended to be an improved version of numpy, specically made to work with

dataframes instead of simple arrays. This allows for easier visualisation of the data. It allows to differentiate columns by giving them names, modify default indexes to datetime formats or custom made indexes. We used it to work with our dataset and help with the cleaning and preprocessing part.

- **matplotlib.pyplot:** Matplotlib is a library that is vastly used to produce graph for Data Visualisation. The pyplot extension is specific to pyton implementation of the library.
- **seaborn**: Seaborn is a graphic enhancement library that works in a complementary fashion with the matplotlib library, it allows for better treatment of certain types of graphs (such as heat maps)

B. Preprocessing the dataset

Data:

Below we can observe the dataset that we found following important variables:

1) **"X":** x-axis spatial coordinate within the Montesinho park map: 1 to 9.
2) **"Y":** y-axis spatial coordinate within the Montesinho park map: 2 to 9.
3) **"month":** Indicates the month that fire happened.
4) **"day":** Indicates the day that fire happened.
5) **"FFMC":** index from the FWI system: 18.7 to 96.20
6) **"DMC":** index from the FWI system: 1.1 to 291.3
7) **"DC":** index from the FWI system: 7.9 to 860.6.
8) **"ISI":** index from the FWI system: 0.0 to 56.10.
9) **"temp":** temperature in Celsius degrees: 2.2 to 33.30.
10) **"RH":** humidity in percentage : 15.0 to 100
11) **"wind":** wind speed in km/h: 0.40 to 9.40.
12) **"rain":** outside rain in mm/m2 : 0.0 to 6.4
13) **"area":** the burned area of the forest (in ha): 0.00 to 1090.84 (this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform)

The dataset has also the next characteristics:

Data Set Characteristics:	Multivariate	**Number of Instances:**	517	**Area:**	Physical
Attribute Characteristics:	Real	**Number of Attributes:**	13	**Date Donated**	2008-02-29
Associated Tasks:	Regression	**Missing Values?**	N/A	**Number of Web Hits:**	1019925

Fig. 4. Data characteristics

So in order to work with our data set we needed to clean it and make sure that all data were correct and complete. To do so, we decided to create a python class that would represent a Data Access Objet (DAO).This was intended with the goal to create a class that would later allow us to work with a subset of methods that we created in order to work with our data set in an easier fashion. To keep the code clear, we separated the creation of the class itself from the main file. We created a forestFireDataset.py file that would host our DAO class. Since we worked with a .csv file, we needed to convert it to a more usable form in python. For this matter, we used the Pandas library, we start by gathering the length of the data set which we will later need for many reasons. Creating a class object implied a specific structure that was hard to assess at first but proved very useful in the later steps of the code.

To work on the processing and cleaning of the data, we created a function that would take care of dropping null values to avoid errors later when using the models. We also transformed the categorical data into numerical ones. The reason for this is that in order to work with the models, later on, it was necessary to convert categorical labels into numeric ones for the supervised learning to work efficiently, we replaced the month with values ranging from 1 to 12 and for the day, values from 1 to 7.

We then proceeded to split the data as they would suit different purposes, the first columns are the data describing the fires themselves, which would be needed by the regressors to identify and predict the area burnt by a certain fire. On the other hand, the last columns are the meteorologic data, that would be needed for the classification process to identify and then predict the future occurrences of wildfires.

The next step comes from a decision that we took after an initial test, our data set is composed of 517 entries, when we split the data into the training set and the test sample, we realised that the models struggled to find significant results. This gave us an idea to generate some randomised data that would follow the same distribution (mean and standard deviation). To do so, we created a generate_data function that would use the .random method embedded in NumPy to create a data frame of randomly generated data that could be concatenated into the data set and later implemented in the model training to improve accuracy.

Next, we needed to create dummy variables to "one hot encode" the data, meaning we needed to sort the values into a binary variable for the classifier to be able to identify fire occurrences, 1 represented any value where a fire occurred and the 0 when there wasn't. For categorical data, we used the labelEncoder function from sklearn with the fit_transform method to modify the data for the application of a classifier.

For visualisation purposes, we wanted to show the kernel density distribution for specific features, the function kde_plot uses the seaborn method "kdeplot" to chart the distribution of the values of the target column. We also added categorization labels to the column area in order to make it clearer when charted, this is the goal of the area_categorisation function.

Then we reassembled all those initial functions into one bigger function called eda (Exploratory Data Analysis). The eda function takes our dataset as input and applies multiple subfunctions to it. We plot the "Area" column's KDE to see the distribution of the fire damages, then we create a new column containing the new label categorisation via the area_categorisation function. With this, we can chart the daily and monthly data of the fire damages into a barplot. Then we compute the correlation matrix between all the features and we chart them into a heatmap for visual clarity.

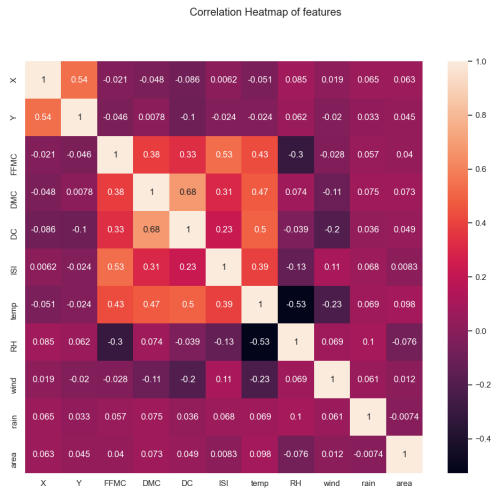

Fig. 5. Correlation matrix of the features

C. Implementation of the supervised learning models

Once the DAO class was completed, we implemented the models and start their training and testing, thus we now switch to the main.py file. First, we imported all the necessary functions from sklearn, this included the models themselves as well as the DAO class that we import from our forestFireDataset file. We declare our dataset as an instance of the class we created and this will now allow us to access all the functions we created earlier to be called as methods of our dataset. We then process our data via the process_data method and plot our data visualisation functions via the exploratory data analysis method (eda). This returns the heatmap, the forest fire damage barplot for daily and monthly data and the distribution of the burnt area for all the recorded fires. After that, we start initialising the parameters for our regression and classification models. We begin with the creation of the kernel for our models, we decided to use the Radial basis function kernel. We then implement an array that will store our different models. We then proceed to fit our models via a for loop with conditions, which aims to attribute the right encoding depending on whether we're dealing with a classifier or a regressor, right after the encoding, we start the training of the model on the split training sample. Finally, we compute the scoring metrics to evaluate the performance of our models. We use the R-Squared(R2) score as well as the Mean absolute error to see which model yields the best results.

V. Results and Conclusion

A. Results

Getting ready the kernel and running the code once we get these results:

— Exploratory Data Analysis —

Feature: area
Skew: 12.8469
Kurtosis: 194.1407

— Gaussian Regression —

R-squared score: 0.7036
Mean Absolute Error: 23.1460

— Ridge Regression —

R-squared score: -0.2360
Mean Absolute Error: 27.2180

— Random Forest Regressor —

R-squared score: 0.8427
Mean Absolute Error: 8.9739

— Support Vector Machine Regressor —

R-squared score: -0.1905
Mean Absolute Error: 25.6024

— Random Forest Classification —
R-squared score: 0.9724
Mean Absolute Error: 24.8556

— Support Vector Machine Classification —
R-squared score: 0.4681
Mean Absolute Error: 26.3195

R-Squared:

To understand what negative value of coefficient of determination R^2, we need to explain the measurement itself. R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit). R-squared can take any values between 0 to 1. Although the statistical measure provides some useful insights regarding the regression model, the user should not rely only on the measure in the assessment of a statistical model. The figure does not disclose information about the causation relationship between the independent and dependent variables. In addition, it does not indicate the correctness of the regression model. Therefore, the user should always draw conclusions about the model by analyzing r-squared together with the other variables in a statistical model.

- R-squared = Explained variation / Total variation and is always between 0 and 100%
- 0% indicates that the model explains none of the variability of the response data around its mean.In other words,

the squared error of your regressor fit is same as the squared error for a fit that always returns the mean of your targets.
- 100% indicates that the model explains all the variability of the response data around its mean. This is not really reliable as the measure can falsely indicate total fitness.

But what happens with a negative R^2?

A negative R^2 means that your regressor fit has a higher squared error than the mean fit. In a few cruel words, it means that it performs worse than the mean fit. Furthermore, R^2 compares the fit of the chosen model with that of a horizontal straight line (the null hypothesis). If the chosen model fits worse than a horizontal line, then R^2 is negative. Note that R^2 is not always the square of anything, so it can have a negative value without violating any rules of math. R^2 is negative only when the chosen model does not follow the trend of the data, so fits worse than a horizontal line. In other words, it simply means that the chosen model (with its constraints) fits the data really poorly.

Mean Absolute Error (MAE)

The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures accuracy for continuous variables. Expressed in words, the MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

B. Conclusion

Based on our results, we can derive some specific conclusions. First of all, we can understand clearly that Ridge Regression and Support Vector Machine Regressor are out of usage, as they give us a negative R-squared. Moreover, gathering the data of the rest regressors and compare them to each other we get a good R-squared score in Gaussian Regressor, Random Forest Regressor and Random Forest Classification. We drop SVM Classification because of its low R-squared measurement. Then, re-checking the rest regressors we find similar MAEs (around 26.0˜), except for Random Forest Regressor, which has Mean Absolute Error: 8.9739. With the lowest deviation and a good R-squared score: 0.8427 we conclude that Random Forest Regressor is the most suitable one to interpret our dataset.

VI. References

- https://stackoverflow.com/
- https://stats.stackexchange.com/
- https://www.datasciencecentral.com/profiles/blogs/regression-analysis-how-do-i-interpret-r-squared-and-assess-the
- https://corporatefinanceinstitute.com/resources/knowledge/other/r-squared/ eumetrain.org/data/4/451/english/msg/ver_cont_var/uos3/uos3_ko1.htm
- https://archive.ics.uci.edu/ml/datasets/forest+fires
- https://www.dw.com/en/portugal-struggles-to-get-forest-fires-under-control/a-55039934
- https://www.nationalgeographic.com/science/article/how-to-live-with-mega-fires-portugal-forests-may-hold-secret
- https://en.wikipedia.org/wiki/June_2017_Portugal_wildfires
- Cortez and Morais, 2007, P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. Available at: [Web Link]
- Igor Drobyshevab, Nina Ryzhkovad, Jonathan Edene, Mara Kitenbergac, Guilherme Pintoh, Henrik Lindbergf, Folmer Krikkeng, Maxim Yermokhini, Yves Bergeronb, Alexander Kryshend Trends and patterns in annually burned forest areas and fire weather across the European boreal zone in the 20th and early 21st centuries]