PREDICTIVE ANALYSIS OF WILDFIRES USING MACHINE LEARNING: A METHODOLOGICAL COMPARISON

Pablo De Ramon[1]*, Sergi García*[2]*, Ivan Quirante*[3]*, Amir Ingher*[4]*, Giang Le*[5]*, Vasco Pearson*[6]

[1]School of Telecommunications Engineering, Universitat Politècnica de Catalunya, Barcelona, Spain
[2]Polytechnic School of Engineering, Universitat Politècnica de Catalunya, Vilanova i la Geltrú, Spain
[3]School of Industrial Engineering, Universitat Politècnica de Catalunya, Barcelona, Spain
[4]Department of Computer Science, Aalto University, Espoo, Finland
[5]Department of Architecture, Technical University of Darmstadt, Darmstadt, Germany
[6]Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
{pablo.de.ramon, sergi.garcia.ibanez, ivan.quirante}@estudiantat.upc.edu, amir.ingher@aalto.fi
huong_giang.le@stud.tu-darmstadt.de, vascopearson@tecnico.ulisboa.pt

ABSTRACT

In the summer of 2022, Europe reached record values in terms of maximum temperatures and number of fires, leading to a corresponding record in fire damage [1][2]. This affected almost every country in the continent, and especially those in the south, like Portugal. According to the Institute for Nature Conservation and Forests (ICNF), Portugal's forest fires affect nowadays a third more area than last decade's average, having burned more than 106.500 hectares this year, and making the country struggle to control this natural hazards. Hence, Portugal has been chosen as a case study to help predict the area burned by forest fires using inexpensive meteorological data and a wide assortment of machine learning techniques. The resulting model of this work has managed to predict the amount of fire burned more accurately than previous literature [3], and we hope that the conclusions obtained will be useful in future forest fire prediction and prevention.

Index Terms— Wildfire Science, Human-Centered Machine Learning, Regression, Support Vector Regression

1. INTRODUCTION

While climate change has been of great concern during the last few decades, in the immediate present society is beginning to experience the brutal consequences scientists had anticipated years ago. Rapidly rising temperatures, longer summers, shorter winters, and heat waves, are harmful byproducts of a rapidly growing population, over-consumption and lack of environmental regulation among many other interrelated factors. Critically, this radical shift in climate conditions has led to a global increase in the number of wildfires (and the amount of land ravaged by them), especially in many European countries, which are the central focus of this paper.

Forest fires are a major environmental issue not only because they create economical and ecological damage, but also because they directly threaten human life. Therefore it is of utmost importance to accurately predict where fires are likely to occur and how much territory will be affected. Properly harnessing Machine Learning techniques can be a powerful tool to achieve this critical task. In particular, we aim to predict the area damaged by a forest fire utilizing real-time and non-costly meteorological data.

The present work is organized as follows: section 1 is a brief introduction of the problem tackled, as well as an outline of the different parts of the paper. In section 2, the main database is presented, and the data points and features are identified, as well as the label to be predicted. In section 3, we discuss the methodology used to process the data (regression methods), the loss functions tested, the hyper-parameters of training algorithms, and finally the model validation technique and benchmarks. In section 4, a comparison of training and validation errors for all models is made in order to choose the best one, which is finally discussed in section 5, where the test-set error is analysed with full transparency and possible improvements are suggested.

2. PROBLEM FORMULATION

In this study we consider forest fire data from the Montesinho natural park, which is located in the Trás-os-Montes north-eastern region of Portugal. The database consists of a total of 517 rows and 13 columns, each data point being a set of environmental conditions collected from a specific area of the park at a certain month and day of the week. The data was collected between January 2000 and December 2003 every time a forest fire occurred. Columns ranging from 1 to 12

Special thanks to Prof. Dr. Alex Jung for his fantastic insight and passionate teaching

represent a certain feature of the data and are shown in table 1.

Attribute	Description
X	x-axis coordinate (from 1 to 9)
Y	y-axis coordinate (from 1 to 9)
month	Month of the year
day	Day of the week
FFMC	Fine Fuel Moisture Code
DMC	Duff Moisture Code
DC	Drought Code
ISI	Initial Spread Index
temp	Outside temperature (ºC)
RH	Outside relative humidity (%)
wind	Outside wind speed (km/h)
rain	Outside rain (mm/mm^2)
area	Total burned area (ha)

Table 1. Data features (adapted from [3])

Also shown in the last row of table 1 is the label or quantity of interest: the total burned area in ha of each incident. Hence, Machine Learning methods for Regression have been implemented due to the numerical characteristic of the dependent variable. The features in this dataset are all numerical with the exception of the month and day, which are categorical features with 12 and 7 categories, respectively. All the numerical features are continuous except for the coordinates which are discrete (values from 1 to 9). FFMC, DMC, DC and ISI are components of the forest Fire Weather Index (FWI), a Canadian system for rating fire danger. FFMC represents the moisture content surface litter and influences ignition of a fire and its 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.

3. METHODS

Before finding the best predictive model, some research was conducted on the data. According to [3], the information used in the experiment was collected from January 2000 to December 2003 and it was built using two sources. Every time a fores fired occurred, several features such as time, date, spatial location within a 9×9 grid of the Montesinho natural park, the type of vegetation involved, the six components of the FWI system (FFMC, DMC, DC, ISI, BUI and FWI) and the local burned area were registered in one of them. Meanwhile, the other database collected the temperature, relative humidity, wind speed and accumulated precipitation in periods of 30 minutes. However, some modifications were made during data gathering: the source's owners added personally the day and month of the records after consulting with the Montesinho fire inspector, and both the BUI and FWI attributes were discarded since they can be obtained from the rest of the variables. Finally, both sources were combined and the resulting database was made up of 517 rows - i.e, data points - and 13 columns.

The dataset contains a total of 247 samples classified as zero area burnt, which correspond to fires that affected an area smaller than 100m^2. For this reason, in addition to the fact that small fires are more frequent, a positive skew is observed if a histogram of the label is plotted. In order to reduce this effect and improve symmetry, a logarithm transformation has been applied to the target attribute: area' $= \ln(\text{area} + 1)$, resulting in the distribution of frequencies depicted in Figure 1.

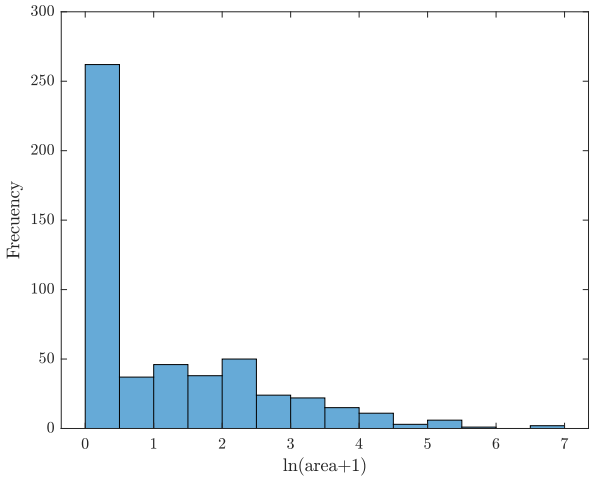

Fig. 1. Histogram of $\ln(\text{area} + 1)$

Finally, a last step has been made during the data preprocessing: the month and the day have been transformed to numbers between ranges 1-12 and 1-7, respectively.

As for the feature selection, the final models use the 12 attributes to predict the amount of area burnt because lower validation errors are obtained when compared with smaller feature subsets.

Since it is a Regression problem, at first two loss functions were considered:

$$\text{Mean Squared Error} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad (1)$$

$$\text{Mean Absolute Error} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \qquad (2)$$

However, MSE was finally chosen because, on the one hand, using a loss function more sensitive to big differences, which usally appear when predicting big fires, would make the model try to fit better these datapoints and focus less on small burned areas that are over-represented in the dataset.

On the other, MSE is a widely-used loss function because it has got appealing characteristics in terms of computational efficiency: it is a convex and differentiable function.

Finally, [3] has been used as a reference document for benchmarking purposes. Their results show that if an error of 1 ha is accepted when testing the final models, 46% of the examples are accurately predicted, whereas this value increases up to 61% if the admissible error is 2 ha. These conclusions have been used as a baseline to evaluate the performance of the final selected model in the test set.

As for the model selection, first a correlation matrix was plotted to find the degree of correlation between the features and the label. All the correlation coefficients were found to be smaller than 0.1 and that is why it was decided to use a wide assortment of models and see which one achieved the smallest validation error. This models, and the rest of techniques have been applied using the *scikit-learn* library for Python. The dataset has been split into two groups with the function `train_test_split()` from `sklearn.model_selection`: one to train and validate each one of the models, and the other to test the final candidate and compare it with benchmark values, whose size corresponds to the 20% of that of all the dataset. After that, in order to reduce overfitting and improve the generalisation capability of the model, a 10-fold cross validation was implemented to the training and validation set. A function named `kfold()` has been created, which is able to return the mean of the squared error of the folds given an specific estimator and the training-validation set. The different methods and the main motivation lying behind their selection are explained below to put the reader in the proper context before showing the results in Section 4: *Linear Regression (LR)*, *Huber Regression (HR)*, *Ridge Regression (RR)*, and *Lasso Regression (RR)*, which use linear models to fit the training data, were tried even though the correlation matrix did not show any clear linear relation between features and labels. The fact that we want to predict a continuous dependent variable from a number of independent variables was the motivation behind trying these regression models, in addition with their simplicity and interpretability. Later, methods with non-linear models were applied to see if they were more appropriate for the data. The *Polynomial Regression (PR)*, *Decision Trees (DT)* (with Naive (Np) and Cost-Complexity (CCp) pruning), *Support Vector Regression (SVR)* and *Artificial Neural Networks (ANN)* were tried, expecting to get the best results with SVR in the same way as [3].

4. RESULTS

Once all the data has been properly preprocessed and randomly split for calculation, all the models were trained and validated. The `kfold()` function has been applied, and for each model the mean of 10 validation error estimates has been computed. The results obtained are summarized in table 2.

Model	AVE	ATE	Comments
LR	2.27	1.82	
HR	2.18	1.9	
RR	2.27	1.17	Best alpha value: $\alpha = 0.001$
LaR	1.89	-	Best alpha value: $\alpha = 1.43$
PR	2.27	1.82	Best degree: d = 1
DT (Np)	**1.83**	**1.81**	min_samples_leaf = 75
DT (CCp)	1.89	1.89	ccp_alpha = 0.058
SVR	**1.85**	**0.25**	$\gamma = 2000, \mathrm{C} = 1.2, \epsilon = 0.362$
ANN	1.95	1.87	4 hidden layers

Table 2. Average validation (AVE) and Training (ATE) Errors for each model

As it can be observed in table 2, the best mean (in **bold**) for the training and validation errors was obtained with the Decision Tree with Naive pruning, closely followed by Support Vector Regression. Since the initial criteria used to compare between the different models was the smallest validation error, the SVR and the Decision Tree were selected as the candidate models to test the test set.

The DT (Np) results were obtained by fitting the hyperparameter `min_samples_leaf` using cross validation on the training-validation set, looking to minimize the squared loss. This pruning works by restricting the tree from growing fully (and therefore overfitting) by simply setting a minimum required amount of data points per terminal node of the tree, hence its "Naive" designation. To achieve the SVR results, the features were firstly scaled using the `sklearn.preprocessing.StandardScaler()` because the SVR model considers distances between observations and these may vary between scaled and non-scaled data. Standardizing the features makes the model more flexible to new values that are not yet seen in the dataset, and in general allows for a higher accuracy [4]. Next, the parameters of the class `sklearn.svm.SVR()` have been tuned until the optimum value has been reached. Following the results obtained by Pablo Cortez and Aníbal Morais in [3], values of $\mathrm{C} = 3$ and $\epsilon = 3 \cdot \hat{\sigma} \cdot \sqrt{\frac{\ln N}{N}}$ have been used, where $\hat{\sigma}$ is the standard deviation of the label vector used for training and validation and N its length. Afterwards, a for-loop has been implemented to find the best $\gamma \in \{2-9, 2-7, 2-5, 2-3, 2-1, 2, 20, 200, 2000\}$.

Finally, the two candidate models have been used to make predictions over the test set. As mentioned in the previous section, the test set was obtained after splitting the data with `train_test_split`, giving it a size corresponding to the 20% of that of the whole dataset. This percentage has been selected following the Pareto principle[1], which is considered a good initial criterion [4]. With DT (Np) we have obtained an MSE on the test set of 2.22, with a prediction accuracy of

[1]The criterion states that for many systems and situations, 80 percent of the output is determined by 20 percent of the input.

39% considering an absolute error under $2ha$. Using SVR, the MSE in the test set has been 2.2, with 69.23% of the samples being accurately predicted if an absolute error of $2ha$ is admitted. It is worth mentioning that this result is relatively higher than that presented in [3].

5. CONCLUSION

In conclusion, the results obtained with the SVR model have been satisfactory. First of all, the training error (0.25) is fairly small, which means that the inner algorithm has properly fit the data points of the training set. Secondly, a slightly higher validation error (1.85) compared with the training one demonstrates that some overfitting might have occurred and the model could be improved in order to get better generalisation characteristics. Thus, we believe that collecting more training data is a promising task that could alleviate this difficulty. Another interesting approach would be to generate more data using data augmentation techniques, for example. The error on the test set differs from the validation error in 0.7 units (MSE = 2.2), which reveals that even though the model is slightly overfitted, it still has the ability to generalize outside the training data. When an absolute error of $1ha$ was admitted on the test set, the percentage of accuracy was 12.5%, which is quite smaller than the 46% achieved in [3]. However, this value increases up to 69.23% when the admissible value changes to 2 ha as mentioned in the previous section, which is a better result than the established benchmark (61%). As can be seen, there is quite a big difference between the two percentages, probably indicating that there exists further room for improvement in the robustness and accuracy of the SVM model.

6. RELATION TO PRIOR WORK

This paper builds upon results obtained in [3] given that now there is more recent literature on state of the art machine learning models and more computing power to achieve better results. Every improvement of fire predictions is invaluable to land managers and owners who are considering preventative maintenance measures and even small improvements can help save lives and resources. It could also increase public awareness of wildfire risk and prompt people to maintain their yards or plan how to evacuate in case of a fire.

Since these are very hot topics in recent years, there has been an increase in literature on predictive analysis [5] and predictive modeling of wildfires [6]. However, the models are only valid in the places where the training data comes from, that is why it is important that there are models achieving accurate results on data from all over the world.

Portugal has one of the highest wildfire risks in the world. However, aside from [3], there is very little literature on predictive analysis of wildfires in the region. We believe its important to encourage and spur research in the community at the same rate as technology and methods are developed and available, and this is the main goal of the present paper.

7. REFERENCES

[1] World Meteorological Organization, "Europe has hottest summer on record: EU copernicus," *WMO*, Sep 2022.

[2] Ashley Kirk, David Blood, and Pablo Gutiérrez, "Europe's record summer of heat and fires – visualised," *The Guardian*, Jul 2022.

[3] Paulo Cortez and Aníbal de Jesus Raimundo Morais, "A data mining approach to predict forest fires using meteorological data," *Associação Portuguesa para a Inteligência Artificial (APPIA)*, 2007.

[4] A. Aylin Tokuç, "Why feature scaling in SVM?," *Baeldung on Computer Science*, Aug 2021.

[5] GeoPlace and London Fire Brigade, "Case study: Predicting and preventing fires using predictive analytics and the UPRN," *GeoPlace*, 2019.

[6] Younes Oulad Sayad, Hajar Mousannif, and Hassan Al Moatassime, "Predictive modeling of wildfires: A new dataset and machine learning approach," *Fire Safety Journal*, vol. 104, pp. 130–146, 2019.