sensors

MDPI

Article

Machine Learning Methods and Synthetic Data Generation to Predict Large Wildfires

Fernando-Juan Pérez-Porras [1], Paula Triviño-Tarradas [1], Carmen Cima-Rodríguez [2], Jose-Emilio Meroño-de-Larriva [1], Alfonso García-Ferrer [1] and Francisco-Javier Mesas-Carrascosa [1,*]

[1] Department of Graphic Engineering and Geomatics, Campus de Rabanales, University of Córdoba, 14071 Córdoba, Spain; o12pepof@uco.es (F.-J.P.-P.); ig2trtap@uco.es (P.T.-T.); ir1melaj@uco.es (J.-E.M.-d.-L.); agferrer@uco.es (A.G.-F.)

[2] Centro de Investigaciones Aplicadas al Desarrollo Agroforestal, Campus de Rabanales, 14071 Córdoba, Spain; ccima@idaf.es

* Correspondence: fjmesas@uco.es

Abstract: Wildfires are becoming more frequent in different parts of the globe, and the ability to predict when and where they will occur is a complex process. Identifying wildfire events with high probability of becoming a large wildfire is an important task for supporting initial attack planning. Different methods, including those that are physics-based, statistical, and based on machine learning (ML) are used in wildfire analysis. Among the whole, those based on machine learning are relatively novel. In addition, because the number of wildfires is much greater than the number of large wildfires, the dataset to be used in a ML model is imbalanced, resulting in overfitting or underfitting the results. In this manuscript, we propose to generate synthetic data from variables of interest together with ML models for the prediction of large wildfires. Specifically, five synthetic data generation methods have been evaluated, and their results are analyzed with four ML methods. The results yield an improvement in the prediction power when synthetic data are used, offering a new method to be taken into account in Decision Support Systems (DSS) when managing wildfires.

Keywords: imbalanced data; burned area; prediction large wildfire; logistic regression; multi-layer perceptron

Citation: Pérez-Porras, F.-J.; Triviño-Tarradas, P.; Cima-Rodríguez, C.; Meroño-de-Larriva, J.-E.; García-Ferrer, A.; Mesas-Carrascosa, F.-J. Machine Learning Methods and Synthetic Data Generation to Predict Large Wildfires. *Sensors* **2021**, *21*, 3694. https://doi.org/10.3390/s21113694

Academic Editors: Assefa M. Melesse, Rosa Lasaponara and Luke Wallace

Received: 19 March 2021
Accepted: 20 May 2021
Published: 26 May 2021

1. Introduction

Forest ecosystems have an inestimable capacity to sequester carbon dioxide (CO_2) through time [1], which is very important to the global carbon budget [2]. Droughts, insects and pathogens, landslides, hurricanes, and fires have negative impacts on natural environments [3]. Among these, wildfires are the greatest hazards to forest development [4]. Wildfires can be caused by either anthropogenic or natural causes. Anthropogenic causes, such as carelessness or arson, accumulate the highest percentage of forest fire origins [5,6], negatively impacting economy and quality of life in both local and regional scales, in addition to the harm to natural environments. The average area burned in the world in the last 16 years is about 340 million hectares [7], with the latest report on forest fires in Europe indicating that 178,000 hectares were burned [8]. In addition, the number of countries that have been harmed in some way by large wildfires is higher than ever before. This damage is further exacerbated by the process of climate change that our planet is experiencing. For example, as the intensity and frequency of drought periods increase, the follow-up impact of wildfires increase, both in fire intensity and frequency [9].

In this context, the prediction, prevention, and management actions for wildfires are crucial. Decision Support Systems (DSS) for wildfires are powerful tools to prevent and manage forest fires by providing data for efficient resource use [10]. These DSSs are based mainly in (a) prediction models of geospatial (topography, land uses, infrastructures, among others), satellite and meteorological data [11,12], (b) thematic maps and risk indexes

of forest fuel and vegetation [13–15], (c) fire propagation and behavior models [16–18] and (d) programs for planning and coordination of fire departments (human force, land, and/or aerial machinery) [19,20]. Several wildfire simulators have been developed, which combine algorithms with theoretical forest fire spread. These simulations aid in the prediction of fire spread, which in turn allows for more accurate and effective extinguishing plans. However, wildfire behavior is a complex process involving physical and chemical factors. Multiple models have been developed to explain fire behavior [21–26] and multiple simulation algorithms have been developed using Huygens' principle of wave propagation [27,28], infiltration theory [29], fractal theory [30,31], or cellular automata [32–34]. This mathematical modelling of fire is integrated into tools for the simulation of the spread and/or management of wildfires such as FARSITE (Fire Area Simulator-Model Development and Evaluation) [35], Prometheus [17,36], BushFire [37], Fire! [38], FireMaster [39], FireStation [40], or SiroFire [41]. However, there are errors related to the spatial and temporal spread of fire, which are linked to large-scale wildfires [42]. These errors are related to the accuracy of input data and lack of clarity over the relevance of parameters used by algorithms [43].

Wildfire behavior models can be classified with physics-based, statistical, and machine learning (ML) methods. Physics-based methods implement equations of canopy biomass, heat transfer, and fluid mechanics, modeling fire behavior in spatial and temporal dimensions [44,45]. These models demand detailed datasets such as the location and dimensions of trees or fuel mass, which is difficult to register on a large scale [46]. Statistical methods provide adequate models for large areas using different methods such as Poisson regression [47], multiple linear regression [48], logistic regression [49], or Monte Carlo simulation [50] by using data at different scales and resolutions [51]. However, results are not accurate because wildfire spread is a complex and non-linear process. Finally, machine learning methods have been explored, yielding better results, in general, than statistical methods [52]. Among the techniques explored in this group are Random Forest [53], Kernel logistic regression [54], or Artificial Neuronal Networks [55]. Normally, these methods offer better results than those based on statistical models because factors such as temperature, wind, rainfall, slope, orientation or land use interact in multiple and complex forms [56].

While it is necessary to have adequate fire management policies, the majority are focused on fire extinction, with the objective to suppress fire at any cost [57], and they do not address socio-economic and land management issues at the initial phase and spread of wildfire [58]. This strategy depends on budget allocation, and therefore, it has economic limitations. As result, effectiveness to suppress and control a wildfire becomes a problem of the initial attack. It is understood that the best fire-fighting response to control wildfires is restricting their ability to become a large wildfire event. As result, this initial attack is linked to the number and type of resources used at the onset of fire detection [59]. Normally, the initial attack consists of a few fire engines and ground crews, occasionally assisted by aerial vehicles. As a result, fire control is more successful in low wildfire intensity [60]. The number and type of resources used in the initial attack is defined by taking fire risk and fuel moisture indices into account [61–64]. Initial attack success depends on multiple factors such as weather conditions, early detection, or fire services arrival time [65]. In previous research, initial attack has been addressed using scenario-based models [66], mixed and linear [67], or two-stage stochastic models [68] with the objective of determining optimal fire engine and crew dimensions without taking into account geospatial data [69].

Given this scenario, there is no certainty in the success of initial attack actions. In addition, it should be noted that not all wildfires should be suppressed in this manner [70]. Therefore, it is necessary to identify wildfire events with high probability of becoming large wildfires. Factors driving large wildfires vary in time and space, but it is not clear which factors are best to be used in a predictive model [71]. Previous research has been carried out on this [72,73]; however, there is still uncertainty predicting large wildfires [74]. The objective of this study was to determine a methodology to predict, at the first moment

of wildfire detection, if it is going to become a large wildfire using ML techniques to support initial attack planning.

2. Materials and Methods

2.1. Study Area

The study area is located in southwestern Spain (37°28′42″ N, 6°54′19″ W, WGS-84), more specifically in western Andalusia, covering the province of Huelva. Every year, there are large recurrent wildfires in the area, defined, in this study, as those that exceed a burn area higher than 500 hectares [75,76]. While in Spain as a whole, the percentage of large wildfires is low, 0.48% in 2017, in Huelva, there is always at least one per year [8]. As an example, in 2018, there were only three large wildfires in Spain, one of which was in the study area [8]. This makes this area of special interest when modeling and predicting the presence of large wildfires. Figure 1a shows the location of 210 wildfires lasting more than 6 h within the study area between 2000 and 2018. The average burnt area surface was equal to 637 hectares, with a maximum equal to 34,290 hectares, which occurred on 27 July 2004. As Figure 1b shows, the number of large wildfires were significantly smaller than normal wildfires. This imbalanced sample makes the results using machine learning techniques biased toward the majority class. For this reason, this study analyzes different sample balancing techniques through the generation of synthetic data.

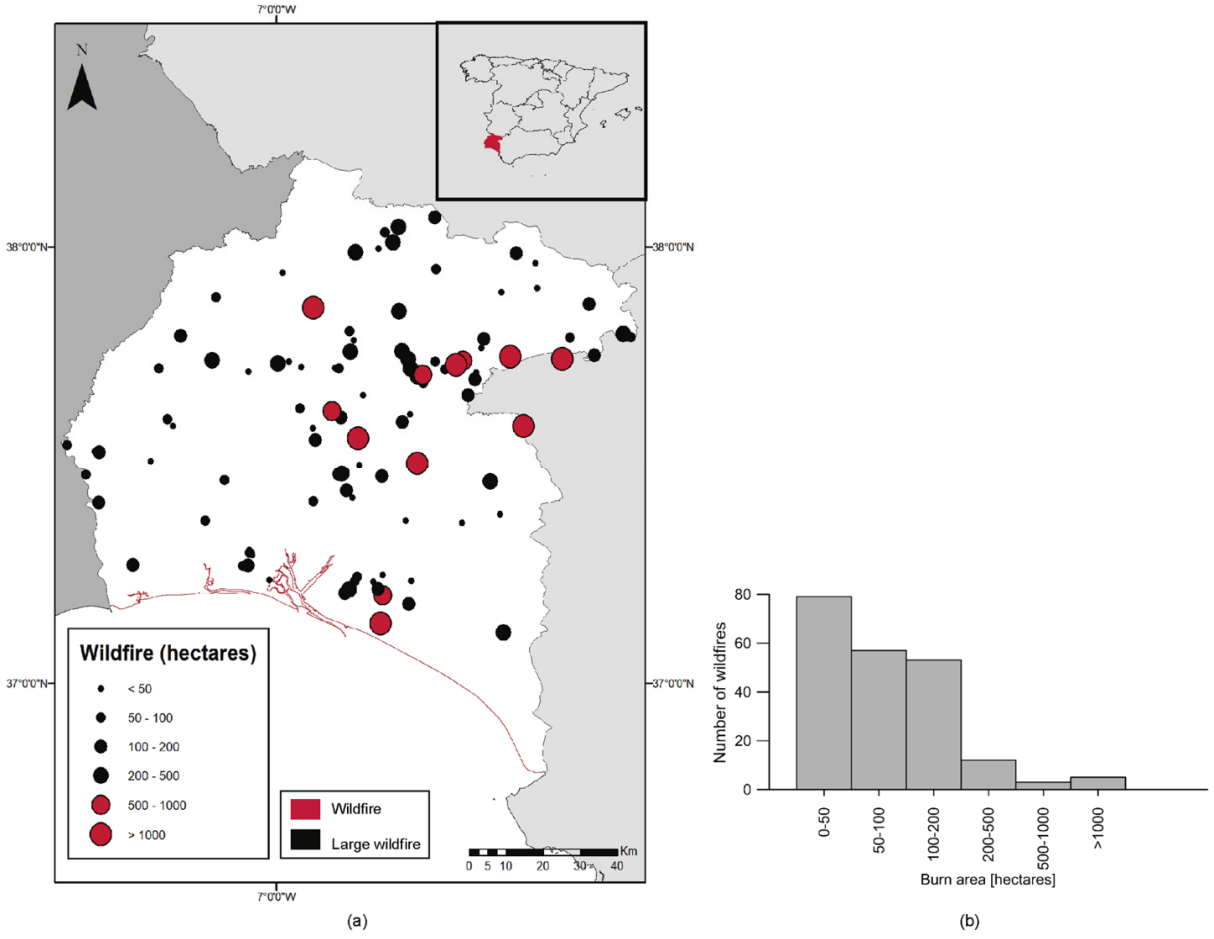

Figure 1. Distribution of wildfires: (**a**) Location of wildfires within study area between 2000 and 2018, and (**b**) histogram of frequency of wildfires by area.

2.2. Data Analysis

A total of 20 variables were analyzed for predicting the occurrence of large wildfires, including meteorological and environmental data as well as data calculated from Landsat and Moderate Resolution Imaging Spectroradiometer (MODIS) scenes [77] (Table 1). This variable selection was based on previous research [78–80]. The environmental variables were obtained from the Environmental Information Network of Andalusia (REDIAM) [81]. These variables have an annual–temporal resolution that is far too involved to describe in this paper; however, the details of and process of obtaining these variables can be found at http://www.juntadeandalucia.es/medioambiente/site/rediam (accessed on 17 December 2020). For the meteorological variables, two data sources were used. Firstly, data from the Spanish Meteorological Agency (AEMET) [82] were used to characterize pre-existing variables of the total study area before wildfire incidences, while on-site data provided by the Regional Forest Fire Fighting Plan of Andalusia (INFOCA) [83] were used to define the meteorological conditions of specific wildfire locations. The variables in the study area characterize the seasonal behavior of the year in which wildfires occur. Thus, the following variables were used:

- Burn area mask: binary raster coverage where a value equal to 1 has been assigned to each pixel that has been affected by a wildfire and a value equal to 0 has been assigned to those pixels that have not been affected by a wildfire.
- Normalized Difference Vegetation Index (NDVI) [84]: calculated from the closest Landsat scene for all burn areas, obtained from the Google Earth Engine.
- Land Surface Temperature: obtained from the closest MOD11A1.006 [79], from the MODIS program, for all burn areas, obtained from the Google Earth Engine.
- Fuel model: fuel modeling adapted to the characteristics of the Mediterranean landscape [85].
- Danger index: reflects the probability of wildfire spread once started.
- Fuel model risk index: includes information on forest fuel patterns as well as on those areas that, despite not having forest vegetation, have agricultural crops susceptible to fire spread.
- Watershed's fuel model risk: represents the mean value of fuel model risk index at watershed scale.
- Geographical risk: measured by slope, annual hours of insolation, and density of drainage points
- Watershed's surface risk: the mean value of geographical risk at watershed scale.
- Local risk index: refers to the risk of affecting human or heritage elements due to their exposure to wildfire, taking into account urban areas, transport networks, elements of historical heritage, power lines, and fuel pipelines.
- Watershed's meteorological risk index: shows at watershed scale those factors related to meteorological conditions that will influence the development of the wildfire. These factors are wind speed and vegetation water deficit.
- Water stress risk index: represents humidity state of vegetation based of the relation between rainfall and evapotranspiration.
- Historical risk index: shows the probability of a wildfire occurring as a function of the historical frequency of wildfires.
- Slope risk index: measures the influence on wildfire behavior of slope as it favors the vertical continuity of fuel.
- Forest vulnerability risk: evaluation the quality of forest ecosystems, measuring continuity in terms of total area of tree coverage.
- Wind speed, wind direction, relative humidity, and mean temperature: measured in real time by a weather station site close to a wildfire.
- Water stress: measures humidity content of soil.

Table 1. Variables analyzed in the prediction of the occurrence of a large fire.

Variables	Sub-Variable	Type	Source	Temporal Resolution
Burn area mask		Binary categorical	INFOCA	Annual
NDVI		Numerical	Landsat	Near real time
Land Surface Temperature		Numerical	Modis	Near real time
Fuel model		Categorical	REDIAM	Annual
Danger index	Risk 1	Numerical	REDIAM	Annual
Watershed's fuel model risk	Risk 2	Categorical	REDIAM	Annual
Watershed's surface risk		Categorical	REDIAM	Annual
Local risk index		Categorical	REDIAM	Annual
Wathersed's meteorological risk index		Categorical	REDIAM	Annual
Fuel model risk index		Categorical	REDIAM	Annual
Water stress risk index		Categorical	REDIAM	Annual
Historical risk index		Categorical	REDIAM	Annual
Slope risk index	Risk 3	Categorical	REDIAM	Annual
Geographical risk		Cateogorical	REDIAM	Annual
Forest vulnerability risk		Numerical	REDIAM	Annual
Wind speed		Numerical	INFOCA	Real-time
Wind direction		Categorical	INFOCA	Real-time
Water stress		Numerical	AEMET	Near real-time
Relative humidity		Continuous	INFOCA	Real-time
Mean temperature		Continuous	INFOCA	Real-time

Figure 2 summarizes the workflow for classifying wildfire size. Firstly, data were pre-processed to generate new variables (Risks 1, 2, and 3). From this new dataset, the Random Forest (RF) technique was used for identifying significant variables. Since large wildfires are fewer than those belonging to the general class of wildfires, several techniques to create synthetic datasets were analyzed to balance the sample size of both classes. These new datasets, together with the original data, were used by four types of ML classifiers in order to predict the size of a wildfire. Finally, these results were evaluated to detect which type of synthetic data generation method and prediction model provided the best results. The prediction accuracy of both classes, wildfire and large wildfire, were also analyzed based on omission and commission errors.

The high number of candidate predictor variables and the low number of observations can impact the machine learning results [86], and therefore, the accuracy of the classifier can be overly optimistic, resulting in an overfitted model [87]. To this end, all the risk variables from REDIAM were summarized using three mean risk variables (Risks 1, 2, and 3). First, the variables were grouped according to whether they represented danger (Risk 1), risk associated with the individual watersheds (Risk 2), and risk associated with geography (Risk 3). In the case of Risk 1, only one variable was linked, and the variable was renamed. In the case of the variables within Risks 2 and 3, which were discussed previously using REDIAM, they were calculated by the mean value of the sub-variables for Risks 2 and 3. Then, RF was applied to this new set of variables to determine their individual importance. Variable importance was evaluated through the Gini Index and Out of Bag accuracy, measuring the degree of association between a given variable and the classification result [88].

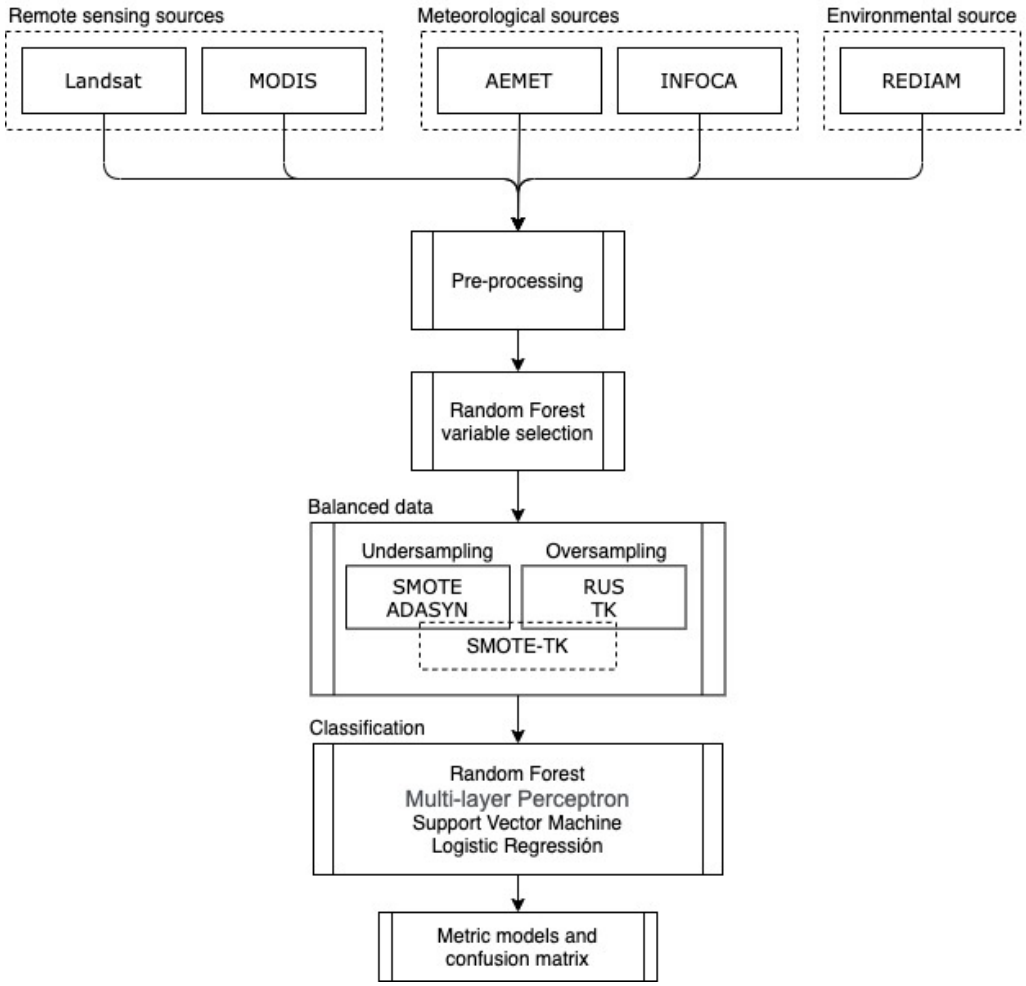

Figure 2. Flowchart used for predicting large fires.

Since the number of large wildfires (sample size equal to 53) was significantly smaller than the totality of wildfires (sample size equal to 157), data were highly imbalanced with the results being biased toward majority class wildfire. As the classifier model assumes, data are drawn from a balanced distribution, and thus, in this case, they produce undesirable results, which can be resolved by balancing techniques [89] divided into two groups: undersampling and oversampling. The former removes data in the majority class, while the latter generates synthetic data in the minority class to balance the ratio between the two classes. For undersampling methods, Random UnderSampling (RUS) and TomeK link (TK) were used. The RUS algorithm balances the classes through random elimination of instances from the majority class [90], while TK detects pairs of samples of nearest neighbors belonging to different classes [91]. TK can either be used, as in this manuscript, in undersampling (majority samples are removed) or cleaning (both samples are removed) mode [92]. The oversampling methods used were Synthetic Minority Oversampling TEchnique (SMOTE) and ADAptative SYNthetic sampling (ADASYN). With the SMOTE algorithm, the minority class is oversampled by forming convex combinations of neighboring samples [93,94], while ADASYM weighs minority samples according to their level of difficulty of learning [95]. Finally, the Synthetic Minority Oversampling Technique-TomeK link (SMOTE-TK) method was used for balancing. This algorithm applies TK as an undersampling technique on the samples that are generated by SMOTE [96].

Next, four different ML classification algorithms were applied to predict wildfire size: Random Forest (RF) [97], Multi-Layer Perceptron (MLP) [98], Support Vector Machine (SVC) [99], and LOGistic regression (LOG) [100]. A grid search was performed for each clas-

sifier to find optimal hyperparameters using Scikit-learn in Python (using GridSearchCV library), as summarized in Table 2. Of the total number of wildfires, 70% were used in the training phase and 30% were used in testing. Training and testing processes have been performed on a virtual machine on Google Colaboratory, which is a free cloud service from Google for machine learning applications, with a Central Processing Unit Intel Xeon 2.30 GHz and a Graphical Processing Unit Tesla K80. Training took less than 9 min and testing only took a few seconds with this configuration.

Table 2. Optimized hyperparameters per classifier and dataset.

Classifier	Hyperparameter	Original	SMOTE	ADASYN	SMOTE TK	TK	RUS
Random Forest	Criterion	Gini	Gini	Entropy	Entropy	Entropy	Gini
	Maximum number of features	4	7	6	7	4	4
	Maximum of level in tree	Automatic	SQRT	Log2	Log2	SQRT	SQRT
	Number of tress	200	200	200	500	500	200
Multi-layer Perceptron	Activation	Identity	Logistic	Logistic	Logistic	Identity	Identity
	Alpha	0.005	0.005	0.01	0.005	0.0001	0.0001
	Hidden layer sizes	50	50	50	50	100	10
	Learning rate	Adaptative	Constant	Constant	Adaptative	Constant	Constant
	Solver	SGD	LBFGS	LBFGS	LBFGS	SGD	SGD
Support Vector Machine	Regularization parameter	100	100	1000	10	100	10
	Gamma	0.0001	0.0001	0.001	0.001	0.0001	0.0001
	Coefficient kernel	RBF	RBF	RBF	RBF	RBF	RBF
Logistic Regression	Inverse of regularization parameter	1	1	1	1	1	1
	Penalty	l2	l2	l2	l2	l2	l2

For the assessment of the ML models, True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) were counted in order to calculate accuracy (1), precision (2), recall (3), specificity (4), Geometric-mean (G-mean) (5), and F1-score (6):

- Accuracy: the ratio of correctly predicted observations to the total observations.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

- Precision: the ratio of correctly predicted positive observations to the total predicted positive observations.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

- Recall: measures how well the classifier can detect positive observations.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

- Specificity: measures how well the classifier can detect negative observations.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (4)$$

- G-mean: equal to the geometric mean of recall and specificity, this shows the balance between classification on the majority and minority class.

$$\text{G} - \text{mean} = \sqrt{\text{Recall} \cdot \text{Specificity}} \quad (5)$$

- F1-score: the harmonic average of precision and recall.

In addition, omission and commission errors were calculated to analyze the results per wildfire class.

$$\text{F1} - \text{score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (6)$$

3. Results

As in the previous analysis, the correlation coefficient between all paired variables used in this study is shown in Figure 3, where positive correlation is represented in blue and negative correlation is represented in red. In addition, color intensity and the circle size are proportional to the correlation coefficients. Wildfire size showed significant correlation with wind speed, LST, mean temperature, risks 1, 2, and 3, forest vulnerability, and relative humidity. In addition, several fuel models were created for wildfire prediction, taking into account plant characteristics and their influence on speed and intensity of flame propagation, as proposed by Rothermel [101]. However, the fuel model variable did not indicate any relationship with wildfire size or the other variables. This explains the lack of classification or mapping of the fuel model in this study. Similar results have been found in other research projects in Spain [102].

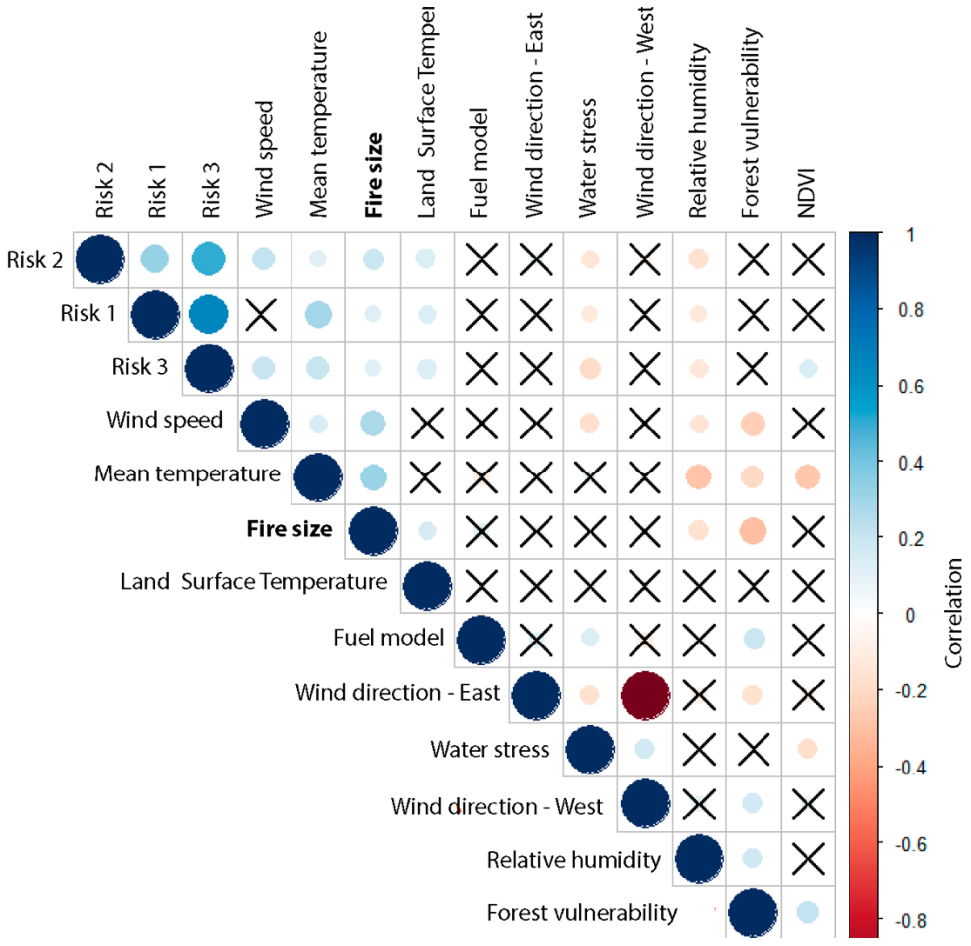

Figure 3. Correlation matrix of variables (significance level for Pearson lower than 0.05; X not significant).

The Gini importance index results with the potential predictor variables are shown in Figure 4a. Wind speed and mean temperature were the most important variables, while risk 1 and risk 3 were the least important. On the other hand, Figure 4b shows that Out of Bag accuracy performs best with four variables. Based on these results, wind speed,

mean temperature, relative humidity, and NDVI were the four selected predictor variables in this study.

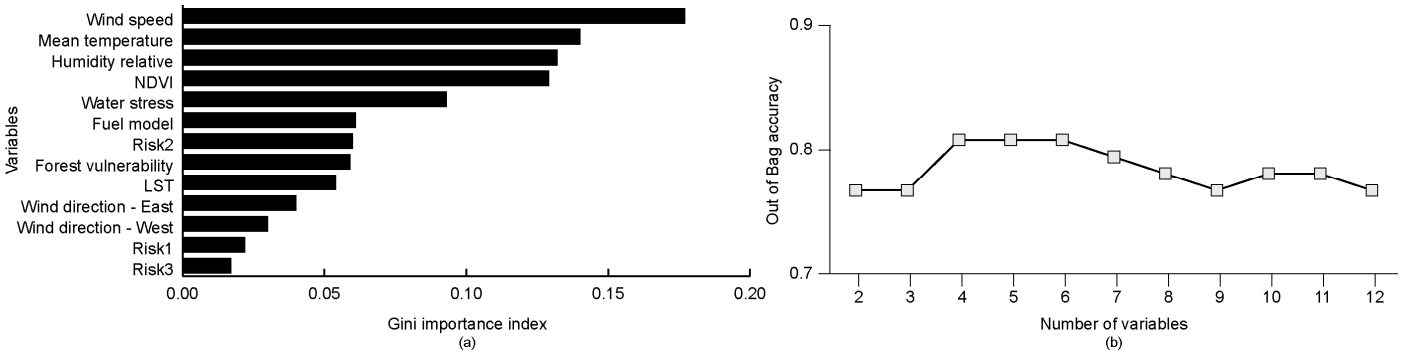

Figure 4. Selection of variables based on (**a**) Gini Index and (**b**) Out of Bag accuracy results.

Once the variables were selected, different synthetic data generation methods were applied to balance the sample. Table 3 shows the sample size of each dataset generated.

Table 3. Sample size of wildfire and large wildfire per original and synthetic datasets.

Dataset	Wildfire	Large Wildfire
Original	189	21
SMOTE	189	189
ADASYN	189	189
RUS	52	52
TK	134	48
SMOTE-TK	171	172

Figure 5 shows a three-dimensional training sample of wildfires (red star) and large wildfires (green circle) throughout a PCA analysis of selected variables for each undersampling and oversampling methods (Figure 5b,f) applied and original data (Figure 5a). On the other hand, Figure 6 shows principal components per dataset. For the undersampling methods, PCA-SMOTE (Figure 6b) generated homogenous distributed synthetic data around the large wildfire region. This region increased with PCA-ADSAYN (Figure 6c). For the oversampling methods, PCA-RUS (Figure 6d) removed data from the wildfire majority class, showing greater differentiation between both classes compared to TK-PCA (Figure 6e). Finally, in Smote-TK-PCA (Figure 6f), wildfires and large wildfires were well differentiated but not as well as in PCA-SMOTE and ADASYN.

Alongside the original dataset, the resulting quality of wildfire size prediction using RF, MLP, Log, and SVC models throughout Recall, F1-score, and G-means are shown in Table 4 and Figure 7. Based on the Recall parameter, the Log model showed the best results. In addition, RUS, SMOTE TOMEK, SMOTE, and ADASYN were the best methods for generating synthetic data for this model, while TOMEK LINKS showed the worst result, yielding similar recall values to the original data using SVC. On the other hand, MLP using SMOTE and SMOTE TOMEK data yielded the best results in the F1-score. As before, the original data gave the worst results. The same results appear with the G-means parameter. The original unbalanced data alone did not provide better results than those described above, therefore showing the advantage of using synthetic data in order to improve wildfire size prediction.

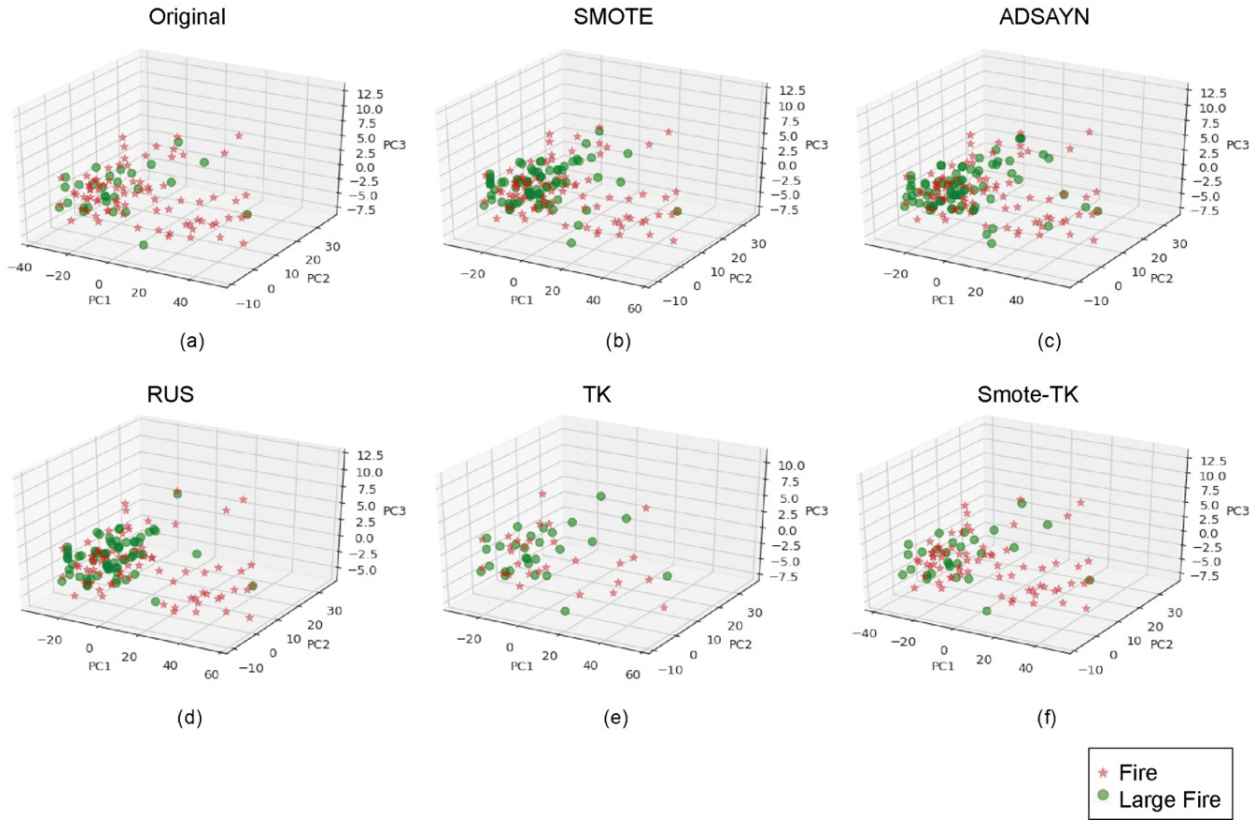

Figure 5. Three-dimensional PCA results per original and synthetic datasets: (**a**) Original data, (**b**) SMOTE, (**c**) ADASYN, (**d**) RUS, (**e**) TK and (**f**) Smote-TK.

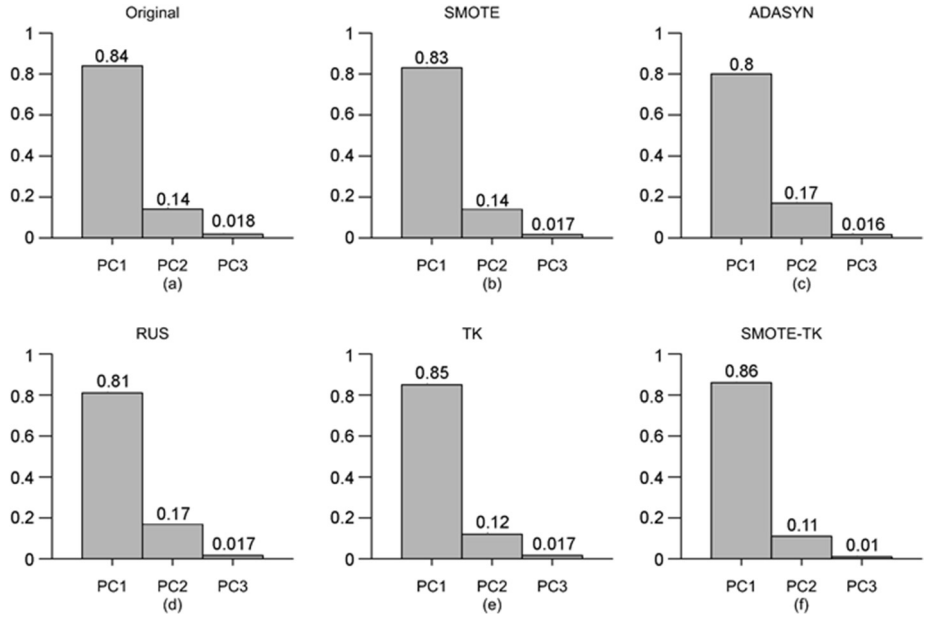

Figure 6. Principal components ranked by variance per original and synthetic datasets: (**a**) Original data, (**b**) SMOTE, (**c**) ADASYN, (**d**) RUS, (**e**) TK and (**f**) Smote-TK.

Table 4. Accuracy and 95% confidence interval of wildfire size prediction based on (a) Recall, (b) F1-score, and (c) G-means.

Accuracy Parameter	Classifier	Random Forest	Multi-Layer Preceptron	Support Vector Machine	Logistic Regression
Recall	Original	0.37 (0.32–0.75)	0.25 (0–0.70)	0.37 (0–0.61)	0.12 (0.04–0.60)
	SMOTE	0.62 (0.61–0.96)	0.75 (0.63–0.95)	0.62 (0.41–0.91)	0.87 (0.49–0.91)
	ADASYN	0.62 (0.61–0.97)	0.50 (0.45–0.95)	0.62 (0.56–0.96)	0.87 (0.43–0.92)
	SMOTE TOMEK	0.70 (0.72–0.99)	0.75 (0.66–0.95)	0.62 (0.48–0.93)	0.87 (0.52–0.90)
	TOMEK LINKS	0.5 (0.41–0.90)	0.52 (0.11–0.88)	0.37 (0.09–0.69)	0.51 (0.14–0.71)
	RUS	0.65 (0.49–1)	0.62 (0.42–0.97)	0.62 (0.15–0.98)	0.87 (0.32–0.97)
F1-Score	Original	0.42 (0.36–0.75)	0.33 (0.05–0.63)	0.42 (0–0.62)	0.18 (0.07–0.62)
	SMOTE	0.5 (0.45–0.75)	0.6 (0.58–0.88)	0.43 (0.42–0.76)	0.53 (0.51–0.75)
	ADASYN	0.5 (0.46–0.91)	0.42 (0.41–0.89)	0.52 (0.48–0.88)	0.53 (0.52–0.77)
	SMOTE TOMEK	0.57 (0.45–0.92)	0.6 (0.42–0.90)	0.47 (0.45–0.79)	0.53 (0.50–0.79)
	TOMEK LINKS	0.57 (0.48–0.82)	0.53 (0.22–0.77)	0.4 (0.16–0.75)	0.47 (0.23–0.74)
	RUS	0.43 (0.43–0.86)	0.29 (0.29–0.84)	0.45 (0.25–0.77)	0.53 (0.38–0.77)
G-means	Original	0.57 (0.38–0.80)	0.48 (0.10–0.72)	0.57 (0–0.71)	0.33 (0–0.58)
	SMOTE	0.67 (0.48–0.85)	0.75 (0.57–0.87)	0.61 (0.43–0.85)	0.75 (0.5–0.90)
	ADASYN	0.67 (0.55–0.92)	0.6 (0.49–0.89)	0.68 (0.51–0.92)	0.73 (0.51–0.82)
	SMOTE TOMEK	0.73 (0.61–0.95)	0.75 (0.59–0.93)	0.65 (0.48–0.82)	0.76 (0.5–0.79)
	TOMEK LINKS	0.67 (0.41–0.81)	0.66 (0.19–0.90)	0.56 (0.21–0.72)	0.63 (0.17–0.70)
	RUS	0.62 (0.45–0.98)	0.57 (0.35–0.91)	0.63 (0.20–0.85)	0.72 (0.34–0.87)

The overall prediction results above were discussed without considering the errors obtained in the analysis of the two wildfire classes. Figure 8 shows the errors of omission and commission for wildfire and large wildfire classes for each model and dataset used. In general, omission errors in the prediction of wildfires (Figure 8(I.a)) were greater than those for large wildfires (Figure 8(I.b)), while on the other hand, the commission error was lower for large wildfires (Figure 8(II.b)) than wildfires overall (Figure 8(II.a)). The lowest omission error in predicting wildfires (Figure 8(I.a)) was obtained when the original data were used regardless of the model applied. However, if synthetic data were used in the same case, the number of wildfires predicted as large wildfire increased. Contrariwise, the omission error in large wildfire prediction decreased if synthetic data were used (Figure 8(I.b)). In this case, the Log model using SMOTE, ADASYN, SMOTE TK, and RUS gave the best results and therefore offered a greater accuracy in predicting large wildfires. On the other hand, the commission error for wildfires (Figure 8(II.a)) was low in those cases where the Log model was applied, using SMOTE, ADASYN, SMOTE TK and RUS synthetic methods. Here, using original data showed the worst results. Finally, MLP using SMOTE and SMOTE TK and RF using TOMEK LINK and original data gave a low commission error (Figure 8(II.b)).

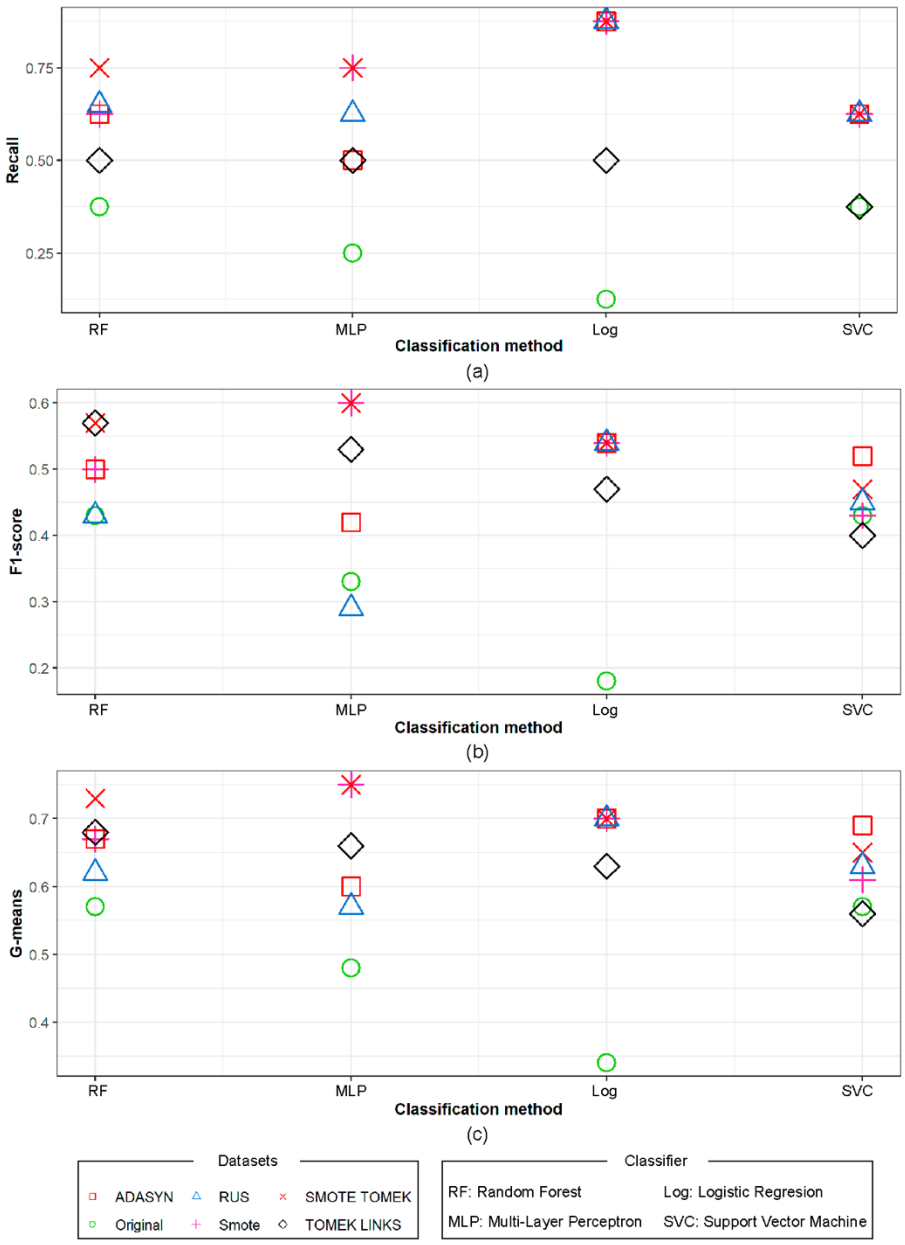

Figure 7. Accuracy of wildfire size prediction based on (**a**) Recall, (**b**) F1-score, and (**c**) G-means.

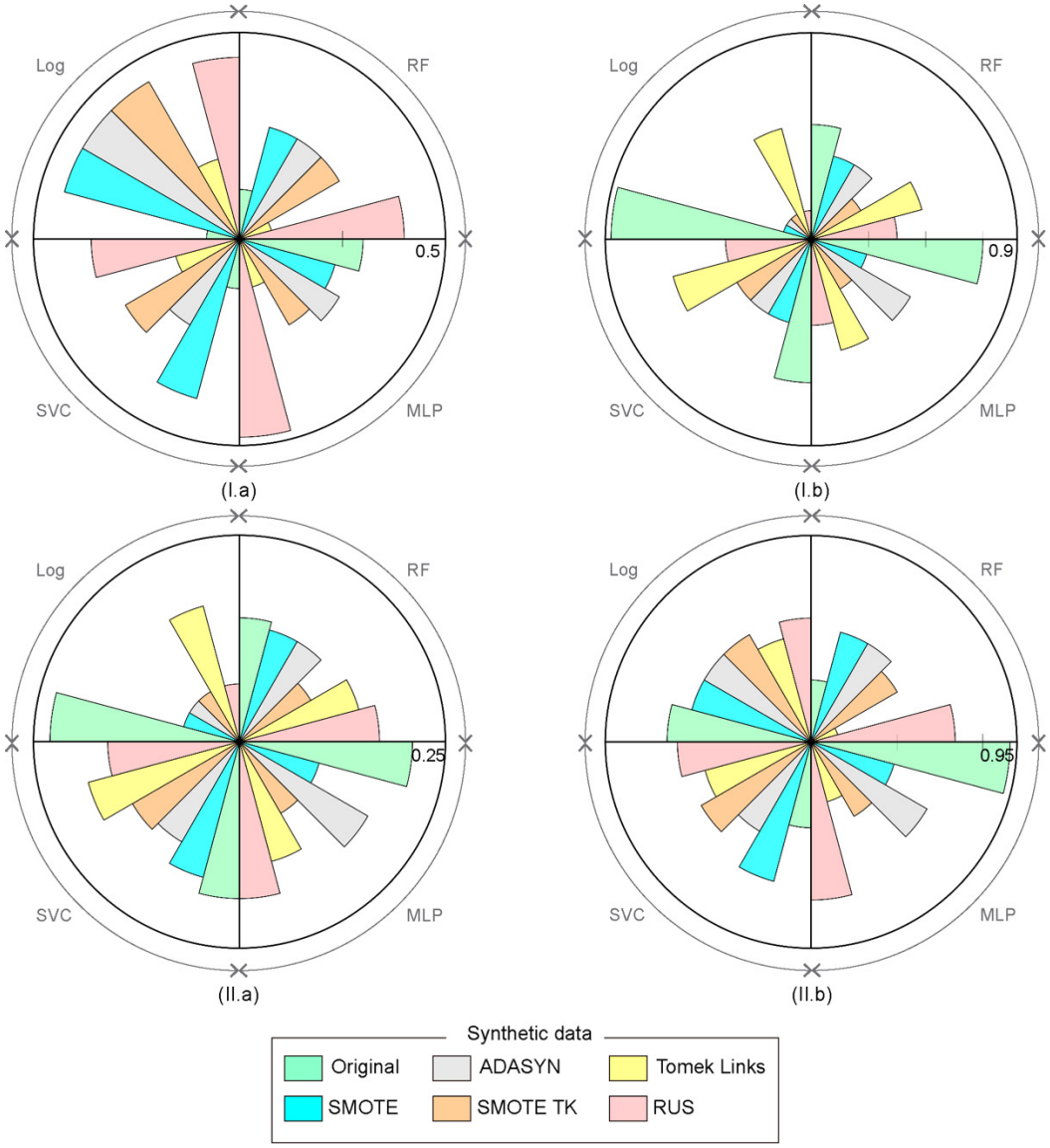

Figure 8. Omission (**I**) and commission (**II**) error for wildfire (**a**) and large wildfire (**b**) classes.

4. Discussion

ML methods applied in studies of burn-area prediction are relatively novel compared to other wildfire applications. To date, these methods are applied to forecast or predict the total area burned and fire occurrence [103,104]. However, these results do not take into account the environmental conditions at the time a wildfire starts. In the proposed methodology, a total of 20 variables have been evaluated to predict the occurrence of a large wildfire. Of these, four have been selected: wind speed, mean temperature, relative humidity, and NDVI. Each of these are linked to real-time or near real-time data. The first three variables coming from a weather station installed close to the wildfire location and the last variable from the most recent Landsat scene at the time of the wildfire, allowing the prediction to be adapted to what is happening at that precise moment of the fire. The selection of meteorological variables is similar to those selected in previous research [74].

The fuel model variable was not selected, although it has been used in previous research projects related to wildfires [105,106]. Our results on fuel models were not conclusive, which was likely due to the low temporal resolution of these data—an aspect that has been detected by other authors working in the same region [102].

Previous studies to model burn area mainly use multiple linear regression models [48,107]; however, there are not many studies using ML techniques in this field [55,108,109]. These studies make predictions of wildfire probability without taking into account the environmental conditions at the onset of the fire. On the other hand, many large wildfire predictions are suboptimal, as they are concentrated in small regions where no general models fit appropriately, which is mostly likely due to small numbers of large wildfires [110]. Therefore, the prediction of large wildfires presents difficulties as they are uncommon events with respect to overall wildfire occurrence. Earlier research, as [111], propose the need to establish a threshold to delimit a large fire event [112]; however, this threshold is debatable [113]. On the other hand, recent ML-based models at continental or global scales for predicting burn areas offer good results in general term but fail to distinguish large wildfires [114]. This imbalance of data justifies the use of synthetic data as proposed in this project.

Based on our results and considering the effectiveness and efficiency, Log and MLP were the best-performing models. In the case of the Log model, it yielded very low errors of omission in the prediction of large wildfires when synthetic data were used, except when using Tomek Links. However, this model was not the most effective, as the wildfire omission error resulted in the highest value of errors. Thus, the model adopts a conservative profile so that the necessary resources will always be mobilized for a large wildfire, assuming that on some occasions, the resources will be oversized. On the other hand, using MLP as a predictive model and SMOTE and SMOTE TK as a technique to generate synthetic data will make the response more efficient but slightly less effective. Thus, the error of omission for a large wildfire is slightly increased, but the error in wildfires will be smaller. Finally, the use of the original data is not recommended, mainly because of the high number of omissions in the prediction of a large fire, indicating the need to balance the sample.

In this study, the use of machine learning applications based on synthetic data to generate a predictive model of the presence of a large wildfire in the early stages has been evaluated. Of all the variables analyzed, the most important were those with a very high temporal resolution rather than historical variables, and therefore, the deployment of sensors over the wildfire area is highly recommended in the initial phase of extinction in order to monitor temperature and wind speed. On the other hand, although the fuel model variable has not been selected in this study, future work should use updated fuel model data to improve the results. Furthermore, we propose the evaluation of this methodology on a large working area, at country or continent scale, to assess its suitability.

5. Conclusions

Wildfires are one of the most dangerous natural hazards across the world and, for this reason, any effort to support its analysis and management is important. Knowing at an early stage whether a wildfire is going to become a large wildfire permits better management of human and material resources. In this study, the use of machine learning methods together with the appropriate selection of variables have provided satisfactory results in the prediction of large wildfires. For this, the selection and processing of data is one of the most important aspects. In this context, the analysis carried out has shown that those data registered at the time of the wildfire were more important than those based on historical series and that it is necessary to balance the data sample due to the higher occurrence of wildfires compared to large wildfires. Given the promising results presented here, the proposed methodology will be applied in future campaigns and will be extended to other regions.

Author Contributions: Conceptualization, F.-J.P.-P.; methodology, F.-J.P.-P., C.C.-R. and F.-J.M.-C.; validation, F.-J.P.-P. and P.T.-T.; formal analysis, F.-J.P.-P. and F.-J.M.-C.; investigation, F.-J.P.-P., C.C.-R. and F.-J.M.-C.; data curation, F.-J.P.-P.; writing—original draft preparation, F.-J.M.-C.; writing—review and editing, F.-J.P.-P. and F.-J.M.-C.; supervision, J.-E.M.-d.-L. and A.G.-F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data generated in this study are available from the corresponding author.

Acknowledgments: Many thanks to the Servicio de Extinción de Incendios Forestales (SEIF) of Plan Infoca in Andalusia (Spain) and the Regional Operational Centre both of which have provided a large amount of meteorological variables in the fires studied.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lü, A.; Tian, H.; Liu, M.; Liu, J.; Melillo, J.M. Spatial and temporal patterns of carbon emissions from forest fires in China from 1950 to 2000. *J. Geophys. Res. Atmos.* **2006**, *111*. [CrossRef]
2. IPCC (Intergovernmental Panel on Climate Change). Climate Change 2007: Synthesis Report. In *Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*; IPCC: Geneve, Switzerland, 2007.
3. Dale, V.H.; Joyce, L.A.; McNulty, S.; Neilson, R.P.; Ayres, M.P.; Flannigan, M.D.; Hanson, P.J.; Irland, L.C.; Lugo, A.E.; Peterson, C.J.; et al. Climate Change and Forest Disturbances: Climate change can affect forests by altering the frequency, intensity, duration, and timing of fire, drought, introduced species, insect and pathogen outbreaks, hurricanes, windstorms, ice storms, or landslides. *Bioscience* **2001**, *51*, 723–734. [CrossRef]
4. Foster, D.R.; Knight, D.H.; Franklin, J.F. Landscape patterns and legacies resulting from large, infrequent forest disturbances. *Ecosystems* **1998**, *1*, 497–510. [CrossRef]
5. Ager, A.A.; Preisler, H.K.; Arca, B.; Spano, D.; Salis, M. Wildfire risk estimation in the Mediterranean area. *Environmetrics* **2014**, *25*, 384–396. [CrossRef]
6. Dimitrakopoulos, A.P.; Bemmerzouk, A.M.; Mitsopoulos, I.D. Evaluation of the Canadian fire weather index system in an eastern Mediterranean environment. *Meteorol. Appl.* **2011**, *18*, 83–93. [CrossRef]
7. Giglio, L.; Randerson, J.T.; van der Werf, G.R. Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (GFED4). *J. Geophys. Res. Biogeosci.* **2013**, *118*, 317–328. [CrossRef]
8. San-Miguel-Ayanz, J.; Durrant, T.; Boca, R.; Libertà, G.; Branco, A.; de Rigo, D.; Ferrari, D.; Maianti, P.; Vivancos, T.A.; Oom, D.; et al. *Forest fires in Europe, Middle East and North Africa*; Publications Office of the European Union: Ispra, Italy, 2018.
9. Williams, C.A.; Hanan, N.P.; Neff, J.C.; Scholes, R.J.; Berry, J.A.; Denning, A.S.; Baker, D.F. Africa and the global carbon cycle. *Carbon Balance Manag.* **2007**, *2*, 3. [CrossRef]
10. Donovan, G.H.; Brown, T.C.; Dale, L. Incentives and Wildfire Management in the United States. In *The Economics of Forest Disturbances*; Holmes, T.P., Prestemon, J.P., Eds.; Springer: Dordrecht, The Netherlands, 2008; pp. 323–340.
11. Illera, P.; Fernández, A.; Delgado, J.A. Temporal evolution of the NDVI as an indicator of forest fire danger. *Int. J. Remote Sens.* **1996**, *17*, 1093–1105. [CrossRef]
12. Bonazountas, M.; Kallidromitou, D.; Kassomenos, P.; Passas, N. A decision support system for managing forest fire casualties. *J. Environ. Manag.* **2007**, *84*, 412–418. [CrossRef]
13. Laxmi, K.S.; Shruti, K.; Mahendra, S.N.; Suman, S.; Prem, C.P. Fuzzy AHP for forest fire risk modeling. *Disaster Prev. Manag. Int. J.* **2012**, *21*, 160–171. [CrossRef]
14. Vadrevu, K.P.; Eaturu, A.; Badarinath, K.V.S. Fire risk evaluation using multicriteria analysis—A case study. *Environ. Monit. Assess.* **2010**, *166*, 223–239. [CrossRef] [PubMed]
15. Thompson, M.P.; Calkin, D.E.; Gilbertson-Day, J.W.; Ager, A.A. Advancing effects analysis for integrated, large-scale wildfire risk assessment. *Environ. Monit. Assess.* **2011**, *179*, 217–239. [CrossRef]
16. Yassemi, S.; Dragićević, S.; Schmidt, M. Design and implementation of an integrated GIS-based cellular automata model to characterize forest fire behaviour. *Ecol. Modell.* **2008**, *210*, 71–84. [CrossRef]
17. Tymstra, C.; Bryce, R.W.; Wotton, B.M.; Taylor, S.W.; Armitage, O.B. *Development and Structure of Prometheus: The Canadian Wildland Fire Growth Simulation Model*; Northern Forestry Centre: Edmonton, AB, Canada, 2009.
18. Parisien, M.-A.; Walker, G.R.; Little, J.M.; Simpson, B.N.; Wang, X.; Perrakis, D.D.B. Considerations for modeling burn probability across landscapes with steep environmental gradients: An example from the Columbia Mountains, Canada. *Nat. Hazards* **2013**, *66*, 439–462. [CrossRef]
19. Dimopoulou, M.; Giannikos, I. Towards an integrated framework for forest fire control. *Eur. J. Oper. Res.* **2004**, *152*, 476–486. [CrossRef]
20. Current, J.; O'Kelly, M. Locating Emergency Warning Sirens. *Decis. Sci.* **1992**, *23*, 221–234. [CrossRef]
21. McArthur, A.G. *Weather and Grassland Fire Behaviour Forestry and Timber Bureau Leaflet No. 100*; Department of National Development: Canberra, Australia, 1966. Available online: https://catalogue.nla.gov.au/Record/752731/Details? (accessed on 24 May 2021).

22. Albini, F.A. Spot Fire Distance from Burning Trees: A Predictive Model. In *Intermountain Forest and Range Experiment Station*; Forest Service: Ogden, UT, USA, 1979; Volume 56.
23. Rothermel, R.C. A Mathematical Model for Predicting Fire Spread in Wildland Fuels. In *Intermountain Forest and Range Experiment Station*; Forest Service: Ogden, UT, USA, 1972; Volume 115.
24. Wagner, C.E. Van Prediction of crown fire behavior in two stands of jack pine. *Can. J. For. Res.* **1993**, *23*, 442–449. [CrossRef]
25. Dupuy, J.-L.; Larini, M. Fire spread through a porous forest fuel bed: A radiative and convective model including fire-induced flow effects. *Int. J. Wildl. Fire* **2000**, *9*, 155–172. [CrossRef]
26. Forbes, L.K. A two-dimensional model for large-scale bushfire spread. *J. Aust. Math. Soc. Ser. B. Appl. Math.* **1997**, *39*, 171–194. [CrossRef]
27. Cunbin, L.; Jing, Z.; Baoguo, T.; Ye, Z. Analysis of forest fire spread trend surrounding transmission line based on rothermel model and huygens principle. *Int. J. Multimed. Ubiquitous Eng.* **2014**, *9*, 51–60.
28. Knight, I.; Coleman, J. A Fire Perimeter Expansion Algorithm-Based on Huygens Wavelet Propagation. *Int. J. Wildl. Fire* **1993**, *3*, 73–84. [CrossRef]
29. Croft, P.; Hunter, J.T.; Reid, N. Forgotten fauna: Habitat attributes of long-unburnt open forests and woodlands dictate a rethink of fire management theory and practice. *For. Ecol. Manag.* **2016**, *366*, 166–174. [CrossRef]
30. McAlpine, R.S.; Wotton, B.M. The use of fractal dimension to improve wildland fire perimeter predictions. *Can. J. For. Res.* **1993**, *23*, 1073–1077. [CrossRef]
31. Song, W.; Weicheng, F.; Binghong, W.; Jianjun, Z. Self-organized criticality of forest fire in China. *Ecol. Modell.* **2001**, *145*, 61–68. [CrossRef]
32. Duff, T.J.; Chong, D.M.; Tolhurst, K.G. Using discrete event simulation cellular automata models to determine multi-mode travel times and routes of terrestrial suppression resources to wildland fires. *Eur. J. Oper. Res.* **2015**, *241*, 763–770. [CrossRef]
33. Karafyllidis, I.; Thanailakis, A. A model for predicting forest fire spreading using cellular automata. *Ecol. Modell.* **1997**, *99*, 87–97. [CrossRef]
34. Hernández Encinas, A.; Hernández Encinas, L.; Hoya White, S.; Martín del Rey, A.; Rodríguez Sánchez, G. Simulation of forest fire fronts using cellular automata. *Adv. Eng. Softw.* **2007**, *38*, 372–378. [CrossRef]
35. Finney, M.A. *FARSITE, Fire Area Simulator—Model Development and Evaluation*; US Department of Agriculture, Forest Service, Rocky Mountain Research Station: Ogden, UT, USA, 1998.
36. CWFGM Steering Committee. Prometheus User Manual v. 3.0.1. Canadian Forest Service, 2004. Available online: https://prometheus.io/docs/introduction/overview/ (accessed on 24 May 2021).
37. Johnston, P.; Milne, G.; Klemitz, D. Overview of bushfire spread simulation systems. *BUSHFIRE CRC Proj. B* **2005**, *6*. Available online: https://www.bushfirecrc.com/sites/default/files/managed/resource/uwa_simulators_overview_0.pdf (accessed on 24 May 2021).
38. Green, K.; Finney, M.; Campbell, J.; Weinstein, D.; Landrum, V. Fire! using GIS to predict fire behavior. *J. For.* **1995**, *93*, 21–25.
39. Eklund, P. A distributed spatial architecture for bush fire simulation. *Int. J. Geogr. Inf. Sci.* **2001**, *15*, 363–378. [CrossRef]
40. Lopes, A.M.G.; Cruz, M.G.; Viegas, D.X. FireStation—an integrated software system for the numerical simulation of fire spread on complex topography. *Environ. Model. Softw.* **2002**, *17*, 269–285. [CrossRef]
41. Coleman, J.R.; Sullivan, A.L. A real-time computer application for the prediction of fire spread across the Australian landscape. *Simulation* **1996**, *67*, 230–240. [CrossRef]
42. Finney, M.A.; Cohen, J.D.; McAllister, S.S.; Jolly, W.M. On the need for a theory of wildland fire spread. *Int. J. Wildl. Fire* **2013**, *22*, 25–36. [CrossRef]
43. Yingying, W.H.Z.W.C.; Sanwei, H.E. Fire Spreading Model Based on CA Scope. *Geomatics Inf. Sci. Wuhan Univ.* **2011**, *5*. Available online: https://xueshu.baidu.com/usercenter/paper/show?paperid=d45fabff11c8e81c7610629d29681362&site=xueshu_se&hitarticle=1 (accessed on 24 May 2021).
44. McGrattan, K.B.; Baum, H.R.; Rehm, R.G.; Hamins, A.; Forney, G.P. *Fire Dynamics Simulator—Technical Reference Guide*; National Institute of Standards and Technology, Building and Fire Research: Gaithersburg, MD, USA, 2000. Available online: https://www.nist.gov/publications/fire-dynamics-simulator-technical-reference-guide-sixth-edition (accessed on 24 May 2021).
45. Tien Bui, D.; Bui, Q.-T.; Nguyen, Q.-P.; Pradhan, B.; Nampak, H.; Trinh, P.T. A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area. *Agric. For. Meteorol.* **2017**, *233*, 32–44. [CrossRef]
46. Pimont, F.; Parsons, R.; Rigolot, E.; de Coligny, F.; Dupuy, J.-L.; Dreyfus, P.; Linn, R.R. Modeling fuels and fire effects in 3D: Model description and applications. *Environ. Model. Softw.* **2016**, *80*, 225–244. [CrossRef]
47. Wotton, B.M.; Martell, D.L.; Logan, K.A. Climate Change and People-Caused Forest Fire Occurrence in Ontario. *Clim. Chang.* **2003**, *60*, 275–295. [CrossRef]
48. Oliveira, S.; Oehler, F.; San-Miguel-Ayanz, J.; Camia, A.; Pereira, J.M.C. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. *For. Ecol. Manag.* **2012**, *275*, 117–129. [CrossRef]
49. Pourghasemi, H.R. GIS-based forest fire susceptibility mapping in Iran: A comparison between evidential belief function and binary logistic regression models. *Scand. J. For. Res.* **2016**, *31*, 80–98. [CrossRef]
50. Conedera, M.; Torriani, D.; Neff, C.; Ricotta, C.; Bajocco, S.; Pezzatti, G.B. Using Monte Carlo simulations to estimate relative fire ignition danger in a low-to-medium fire-prone region. *For. Ecol. Manag.* **2011**, *261*, 2179–2187. [CrossRef]

51. Wittenberg, L.; Malkinson, D. Spatio-temporal perspectives of forest fires regimes in a maturing Mediterranean mixed pine landscape. *Eur. J. For. Res.* **2009**, *128*, 297–304. [CrossRef]
52. Bar Massada, A.; Syphard, A.D.; Stewart, S.I.; Radeloff, V.C. Wildfire ignition-distribution modelling: A comparative study in the Huron–Manistee National Forest, Michigan, USA. *Int. J. Wildl. Fire* **2013**, *22*, 174–183. [CrossRef]
53. Arpaci, A.; Malowerschnig, B.; Sass, O.; Vacik, H. Using multi variate data mining techniques for estimating fire susceptibility of Tyrolean forests. *Appl. Geogr.* **2014**, *53*, 258–270. [CrossRef]
54. Tien Bui, D.; Le, K.-T.T.; Nguyen, V.C.; Le, H.D.; Revhaug, I. Tropical forest fire susceptibility mapping at the Cat Ba National Park Area, Hai Phong City, Vietnam, using GIS-based kernel logistic regression. *Remote Sens.* **2016**, *8*, 347. [CrossRef]
55. Satir, O.; Berberoglu, S.; Donmez, C. Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem. *Geomat. Nat. Hazards Risk* **2016**, *7*, 1645–1658. [CrossRef]
56. Ganteaume, A.; Camia, A.; Jappiot, M.; San-Miguel-Ayanz, J.; Long-Fournel, M.; Lampin, C. A Review of the Main Driving Factors of Forest Fire Ignition Over Europe. *Environ. Manag.* **2013**, *51*, 651–662. [CrossRef] [PubMed]
57. Rodrigues, M.; Alcasena, F.; Vega-García, C. Modeling initial attack success of wildfire suppression in Catalonia, Spain. *Sci. Total Environ.* **2019**, *666*, 915–927. [CrossRef] [PubMed]
58. Fernandes, P.A.M. Forest fires in Galicia (Spain): The outcome of unbalanced fire management. *J. For. Econ.* **2008**, *14*, 155–157. [CrossRef]
59. Lee, Y.; Fried, J.S.; Albers, H.J.; Haight, R.G. Deploying initial attack resources for wildfire suppression: Spatial coordination, budget constraints, and capacity constraints. *Can. J. For. Res.* **2012**, *43*, 56–65. [CrossRef]
60. Gill, A.M. Landscape fires as social disasters: An overview of 'the bushfire problem'. *Glob. Environ. Chang. Part B Environ. Hazards* **2005**, *6*, 65–80. [CrossRef]
61. Ziccardi, L.G.; Thiersch, C.R.; Yanai, A.M.; Fearnside, P.M.; Ferreira-Filho, P.J. Forest fire risk indices and zoning of hazardous areas in Sorocaba, São Paulo state, Brazil. *J. For. Res.* **2020**, *31*, 581–590. [CrossRef]
62. You, W.; Lin, L.; Wu, L.; Ji, Z.; Yu, J.; Zhu, J.; Fan, Y.; He, D. Geographical information system-based forest fire risk assessment integrating national forest inventory data and analysis of its spatiotemporal variability. *Ecol. Indic.* **2017**, *77*, 176–184. [CrossRef]
63. Torres, F.T.P.; Romeiro, J.M.N.; Santos, A.C.d.A.; de Oliveira Neto, R.R.; Lima, G.S.; Zanuncio, J.C. Fire danger index efficiency as a function of fuel moisture and fire behavior. *Sci. Total Environ.* **2018**, *631–632*, 1304–1310. [CrossRef]
64. Cawson, J.G.; Duff, T.J.; Tolhurst, K.G.; Baillie, C.C.; Penman, T.D. Fuel moisture in Mountain Ash forests with contrasting fire histories. *For. Ecol. Manag.* **2017**, *400*, 568–577. [CrossRef]
65. Gonzalez-Olabarria, J.R.; Reynolds, K.M.; Larrañaga, A.; Garcia-Gonzalo, J.; Busquets, E.; Pique, M. Strategic and tactical planning to improve suppression efforts against large forest fires in the Catalonia region of Spain. *For. Ecol. Manag.* **2019**, *432*, 612–622. [CrossRef]
66. Haight, R.G.; Fried, J.S. Deploying Wildland Fire Suppression Resources with a Scenario-Based Standard Response Model. *INFOR Inf. Syst. Oper. Res.* **2007**, *45*, 31–39. [CrossRef]
67. van der Merwe, M.; Minas, J.P.; Ozlen, M.; Hearne, J.W. A mixed integer programming approach for asset protection during escaped wildfires. *Can. J. For. Res.* **2014**, *45*, 444–451. [CrossRef]
68. Wei, Y.; Bevers, M.; Belval, E.; Bird, B. A Chance-Constrained Programming Model to Allocate Wildfire Initial Attack Resources for a Fire Season. *For. Sci.* **2014**, *61*, 278–288. [CrossRef]
69. Arienti, M.C.; Cumming, S.G.; Boutin, S. Empirical models of forest fire initial attack success probabilities: The effects of fuels, anthropogenic linear features, fire weather, and management. *Can. J. For. Res.* **2006**, *36*, 3155–3166. [CrossRef]
70. Costafreda Aumedes, S.; Cardil Forradellas, A.; Molina Terrén, D.; Daniel, S.N.; Mavsar, R.; Vega García, C. Analysis of factors influencing deployment of fire suppression resources in Spain using artificial neural networks. *iForest Biogeosci. For.* **2016**, *9*, 138–145. [CrossRef]
71. Barbero, R.; Abatzoglou, J.T.; Steel, E.A.; Larkin, N.K. Modeling very large-fire occurrences over the continental United States from weather and climate forcing. *Environ. Res. Lett.* **2014**, *9*, 124009. [CrossRef]
72. Bermudez, P.Z.; Mendes, J.; Pereira, J.M.C.; Turkman, K.F.; Vasconcelos, M.J.P. Spatial and temporal extremes of wildfire sizes in Portugal (1984–2004). *Int. J. Wildl. fire* **2009**, *18*, 983–991. [CrossRef]
73. Díaz-Avalos, C.; Juan, P.; Serra-Saurina, L. Modeling fire size of wildfires in Castellon (Spain), using spatiotemporal marked point processes. *For. Ecol. Manag.* **2016**, *381*, 360–369. [CrossRef]
74. Joseph, M.B.; Rossi, M.W.; Mietkiewicz, N.P.; Mahood, A.L.; Cattau, M.E.; St. Denis, L.A.; Nagy, R.C.; Iglesias, V.; Abatzoglou, J.T.; Balch, J.K. Spatiotemporal prediction of wildfire size extremes with Bayesian finite sample maxima. *Ecol. Appl.* **2019**, *29*, e01898. [CrossRef]
75. Rodríguez-Carreras, R.; Úbeda, X.; Francos, M.; Marco, C. After the Wildfires: The Processes of Social Learning of Forest Owners' Associations in Central Catalonia, Spain. *Sustainability* **2020**, *12*, 6042. [CrossRef]
76. Sobrino, J.A.; Llorens, R.; Fernández, C.; Fernández-Alonso, J.M.; Vega, J.A. Relationship between Soil Burn Severity in Forest Fires Measured In Situ and through Spectral Indices of Remote Detection. *Forests* **2019**, *10*, 457. [CrossRef]
77. Justice, C.O.; Townshend, J.R.G.; Vermote, E.F.; Masuoka, E.; Wolfe, R.E.; Saleous, N.; Roy, D.P.; Morisette, J.T. An overview of MODIS Land data processing and product status. *Remote Sens. Environ.* **2002**, *83*, 3–15. [CrossRef]
78. Dennison, P.E.; Brewer, S.C.; Arnold, J.D.; Moritz, M.A. Large wildfire trends in the western United States, 1984–2011. *Geophys. Res. Lett.* **2014**, *41*, 2928–2933. [CrossRef]

79. Hudak, A.T.; Morgan, P.; Bobbitt, M.J.; Smith, A.M.S.; Lewis, S.A.; Lentile, L.B.; Robichaud, P.R.; Clark, J.T.; McKinley, R.A. The Relationship of Multispectral Satellite Imagery to Immediate Fire Effects. *Fire Ecol.* **2007**, *3*, 64–90. [CrossRef]
80. Mueller, S.E.; Thode, A.E.; Margolis, E.Q.; Yocom, L.L.; Young, J.D.; Iniguez, J.M. Climate relationships with increasing wildfire in the southwestern US from 1984 to 2015. *For. Ecol. Manag.* **2020**, *460*, 117861. [CrossRef]
81. Junta de Andalucia. REDIAM-Red de Información Ambienta de Andalucía. Available online: http://www.juntadeandalucia.es/medioambiente/site/rediam (accessed on 17 December 2020).
82. Agencia Española de Meterología AEMET. Available online: http://www.aemet.es/es/portada (accessed on 12 April 2021).
83. Consejería de Agricultura, Ganadería, P. y D.S. INFOCA. Available online: http://www.juntadeandalucia.es/medioambiente/site/portalweb/menuitem.220de8226575045b25f09a105510e1ca/?vgnextoid=2076a5f862fa5310VgnVCM1000001325e50aRCRD&vgnextchannel=321cc98d5b40b410VgnVCM2000000624e50aRCRD (accessed on 12 April 2021).
84. Rouse, J.W.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring vegetation systems in the Great Plains with ERTS. *NASA Spec. Publ.* **1974**, *351*, 309.
85. Rodríguez y Silva, F.; Molina, J.R. *Manual Técnico para la Modelización de la Combustibilidad Asociada a los Ecosistemas Forestales Mediterráneos*; University of Córdoba: Córdoba, Spain, 2010.
86. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; 2012; ISBN 0262304325. Available online: https://www.google.com.hk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjvw8yDuObwAhUHMd4KHcvKA5gQFjACegQIAxAD&url=http%3A%2F%2Fnoiselab.ucsd.edu%2FECE228%2FMurphy_Machine_Learning.pdf&usg=AOvVaw0ivnxQoBAr1Kn4BwTBbNxe (accessed on 24 May 2021).
87. Subramanian, J.; Simon, R. Overfitting in prediction models—Is it a problem only in high dimensions? *Contemp. Clin. Trials* **2013**, *36*, 636–641. [CrossRef]
88. Khalilia, M.; Chakraborty, S.; Popescu, M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med. Inform. Decis. Mak.* **2011**, *11*, 51. [CrossRef] [PubMed]
89. Farquad, M.A.H.; Bose, I. Preprocessing unbalanced data using support vector machine. *Decis. Support Syst.* **2012**, *53*, 226–233. [CrossRef]
90. Geapa, B.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor.* **2004**, *6*, 20–29.
91. Tomek, I. Two modifications of CNN. *IEEE Trans. Syst. Man Cybern.* **1976**, *6*, 769–772.
92. Waldner, F.; Chen, Y.; Lawes, R.; Hochman, Z. Needle in a haystack: Mapping rare and infrequent crops using satellite imagery and data balancing methods. *Remote Sens. Environ.* **2019**, *233*, 111375. [CrossRef]
93. Bhagat, R.C.; Patil, S.S. Enhanced SMOTE Algorithm for Classification of Imbalanced Big-Data Using Random Forest. In Proceedings of the 2015 IEEE International Advance Computing Conference (IACC), Bangalore, India, 12–13 June 2015; pp. 403–408.
94. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
95. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–6 June 2008; pp. 1322–1328.
96. Zeng, M.; Zou, B.; Wei, F.; Liu, X.; Wang, L. Effective Prediction of Three Common Diseases by Combining SMOTE with Tomek Links Technique for Imbalanced Medical Data. In Proceedings of the 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), Chongqing, China, 28–29 May 2016; pp. 225–228.
97. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **2005**, *26*, 217–222. [CrossRef]
98. Ghazikhani, A.; Monsefi, R.; Yazdi, H.S. Online neural network model for non-stationary and imbalanced data stream classification. *Int. J. Mach. Learn. Cybern.* **2014**, *5*, 51–62. [CrossRef]
99. Tang, Y.; Zhang, Y.; Chawla, N.V.; Krasser, S. SVMs Modeling for Highly Imbalanced Classification. *IEEE Trans. Syst. Man Cybern. Part B* **2009**, *39*, 281–288. [CrossRef]
100. Williams, D.P.; Myers, V.; Silvious, M.S. Mine Classification With Imbalanced Data. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 528–532. [CrossRef]
101. Duff, T.J.; Keane, R.E.; Penman, T.D.; Tolhurst, K.G. Revisiting Wildland Fire Fuel Quantification Methods: The Challenge of Understanding a Dynamic, Biotic Entity. *Forests* **2017**, *8*, 351. [CrossRef]
102. Vallejo-Villalta, I.; Rodríguez-Navas, E.; Márquez-Pérez, J. Mapping Forest Fire Risk at a Local Scale—A Case Study in Andalusia (Spain). *Environments* **2019**, *6*, 30. [CrossRef]
103. Cheng, T.; Wang, J. Integrated Spatio-temporal Data Mining for Forest Fire Prediction. *Trans. GIS* **2008**, *12*, 591–611. [CrossRef]
104. Mayr, M.J.; Vanselow, K.A.; Samimi, C. Fire regimes at the arid fringe: A 16-year remote sensing perspective (2000–2016) on the controls of fire activity in Namibia from spatial predictive models. *Ecol. Indic.* **2018**, *91*, 324–337. [CrossRef]
105. Cochrane, M.A.; Moran, C.J.; Wimberly, M.C.; Baer, A.D.; Finney, M.A.; Beckendorf, K.L.; Eidenshink, J.; Zhu, Z. Estimation of wildfire size and risk changes due to fuels treatments. *Int. J. Wildl. Fire* **2012**, *21*, 357–367. [CrossRef]
106. Salis, M.; Laconi, M.; Ager, A.A.; Alcasena, F.J.; Arca, B.; Lozano, O.; de Oliveira, A.F.; Spano, D. Evaluating alternative fuel treatment strategies to reduce wildfire losses in a Mediterranean area. *For. Ecol. Manag.* **2016**, *368*, 207–221. [CrossRef]
107. Flannigan, M.D.; Logan, K.A.; Amiro, B.D.; Skinner, W.R.; Stocks, B.J. Future Area Burned in Canada. *Clim. Chang.* **2005**, *72*, 1–16. [CrossRef]

108. Amatulli, G.; Camia, A.; San-Miguel-Ayanz, J. Estimating future burned areas under changing climate in the EU-Mediterranean countries. *Sci. Total Environ.* **2013**, *450–451*, 209–222. [CrossRef]
109. Yang, L.; Dawson, C.W.; Brown, M.R.; Gell, M. Neural network and GA approaches for dwelling fire occurrence prediction. *Knowl. Based Syst.* **2006**, *19*, 213–219. [CrossRef]
110. Trigo, R.M.; Pereira, J.M.C.; Pereira, M.G.; Mota, B.; Calado, T.J.; Dacamara, C.C.; Santo, F.E. Atmospheric conditions associated with the exceptional fire season of 2003 in Portugal. *Int. J. Climatol.* **2006**, *26*, 1741–1757. [CrossRef]
111. Jiang, Y.; Zhuang, Q. Extreme value analysis of wildfires in Canadian boreal forest ecosystems. *Can. J. For. Res.* **2011**, *41*, 1836–1851. [CrossRef]
112. Coles, S.; Bawa, J.; Trenner, L.; Dorazio, P. Classical Extreme Value Theory and Models. In *An introduction to Statistical Modeling of Extreme Values*; Springer: London, UK, 2001; pp. 45–73. ISBN 978-1-84996-874-4.
113. Tedim, F.; Leone, V.; Amraoui, M.; Bouillon, C.; Coughlan, M.R.; Delogu, G.M.; Fernandes, P.M.; Ferreira, C.; McCaffrey, S.; McGee, T.K.; et al. Defining Extreme Wildfire Events: Difficulties, Challenges, and Impacts. *Fire* **2018**, *1*, 9. [CrossRef]
114. Joshi, J.; Sukumar, R. Improving prediction and assessment of global fires using multilayer neural networks. *Sci. Rep.* **2021**, *11*, 3295. [CrossRef]