Decision Analytics Journal 18 (2026) 100674

Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

An integrated resampling and machine learning framework for predictive analytics of large wildfires

Luís Camacho [a,*], Filipe X. Catry [b], Fernando Bacao [a]

[a] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, Lisboa, 1070-312, Portugal
[b] Centre for Applied Ecology "Prof. Baeta Neves" (CEABN-InBIO), School of Agriculture, University of Lisbon, Tapada da Ajuda, Lisboa, 1349-017, Portugal

ARTICLE INFO

Dataset link: https://drive.google.com/drive/folders/1NhAuRV9b0WRxQmUfp06i0lmPNmNPgOh9?usp=sharing

Keywords:
Machine learning
Predictive analytics
Imbalanced regression
Wildfire forecasting
Risk modeling
Data resampling

ABSTRACT

Accurately predicting the final burned area of wildfires at the moment of ignition is a challenging problem with important implications for disaster management and resource allocation. In this study, we apply recently developed resampling techniques, combined with machine learning algorithms, to improve predictive accuracy for rare, high-impact wildfire events. While the study focuses on wildfires, the methodology addresses a broader challenge in decision analytics: generating reliable predictions from imbalanced datasets to support informed managerial decisions. Our results demonstrate that these techniques can enhance the robustness of predictive models for extreme events, offering insights that may be relevant to other domains where accurate forecasts of rare outcomes are critical.

1. Introduction

Wildfires often pose a threat to ecosystems [1], adversely affect public health [2], and result in significant economic losses [3]. It is therefore not surprising that this topic has attracted substantial interest in recent decades across diverse domains of wildfire science, including fuel characterization, fire detection and mapping, fire weather and climate change, fire occurrence, fire behavior prediction, and fire management, among others [4].

Many researchers have employed machine learning methods to advance their work in the aforementioned domains. The following studies serve as illustrative examples: Silva et al. [5] developed a wildfire warning map for a region of the Brazilian Amazon; Kang et al. [6] utilized deep learning to detect forest fires using data from geostationary satellites; Henderson and Chakrabortty [7] used machine learning techniques to predict the probability of fire ignition in Australia; and Paul et al. [8] developed a model to estimate daily smoke exposure levels in Canada.

Many of these tasks are inherently challenging. One significant difficulty is the imbalanced nature of the data. For instance, when predicting the concentration of fine particulate matter ($PM_{2.5}$) to assess public exposure to wildfire smoke, there are typically far more records of days with low $PM_{2.5}$ concentrations than those with high concentrations. This imbalance poses a serious challenge, as machine learning models often perform poorly on underrepresented cases [9]. In this context, these cases correspond to days with elevated $PM_{2.5}$ levels, which are especially concerning due to their potential effects on human health.

Predicting the burned area of a wildfire at the time of its onset is another domain that has attracted significant research attention, where machine learning techniques have been employed to address the task. Furthermore, this is also an area in which learning is primarily conducted using imbalanced datasets. Wildfire datasets provided by official agencies typically contain far more records of small fires than of medium-sized or large fires, when measured in terms of burned area.

The study by Cortez and Morais [10] is among the earliest contributions in this domain and the first to employ the well-known dataset from the Montesinho Natural Park in Portugal. Since then, subsequent studies have aimed to develop models with improved predictive performance or to further explore the characteristics of this dataset. Beyond Portugal, research on burned area forecasting has also drawn on wildfire data from various geographic areas, including Africa and South America [11], the boreal forests of Alaska [12], Canada [13], South Asia [14], and Turkey [15].

Although statistics from the Global Wildfire Information System (GWIS) [16], based on the MCD64A1 burned area product [17], indicate that the global annual burned area has remained relatively stable in recent years, the number of severe wildfires has been rising [18], and

* Corresponding author.
E-mail addresses: d20190051@novaims.unl.pt (L. Camacho), fcatry@isa.ulisboa.pt (F.X. Catry), bacao@novaims.unl.pt (F. Bacao).

https://doi.org/10.1016/j.dajour.2026.100674
Received 2 September 2025; Received in revised form 7 January 2026; Accepted 8 January 2026
Available online 10 January 2026

several studies suggest that wildfire-related problems are likely to intensify under changing climate conditions [19–21]. Moreover, wildfires lead to a range of environmental impacts, including the degradation of water quality [22,23], the loss of biodiversity [24], increased pollutant emissions and deterioration of air quality [25,26], and the loss of forests in protected areas [27], among other effects. Overall, these trends and environmental impacts highlight the need for further research on wildfires.

However, despite their valuable contributions, some studies overlook a critical characteristic of wildfire datasets: their imbalanced distribution. In particular, the dataset contains significantly fewer instances of medium-sized and large fires compared to small ones, making them especially challenging for machine learning models to learn. Consequently, models often perform poorly in predicting large fires, even when their overall performance, as measured by standard evaluation metrics, appears great. Other studies, while acknowledging the issue, often fail to implement effective strategies that specifically enhance learning from the most challenging cases. As a result, the predictive performance of these models with respect to large fires remains compromised.

Our work addresses the challenge of predicting the burned area of wildfires at the time of ignition, focusing on the most difficult cases and leveraging recent advances in imbalanced regression. This subfield of imbalanced learning focuses on regression problems characterized by two key features: (i) the distribution of target values is uneven, with certain ranges exhibiting low density; and (ii) there is particular interest in accurately predicting samples that fall within these low-density regions. This is directly applicable to our problem, as we aim to map a set of predictor variables to a numerical value representing the burned area of a wildfire and historical wildfire data indicate that large fires occur significantly less frequently than small fires, leading to an imbalanced distribution within the continuous target variable.

The results indicate that imbalanced regression techniques can improve prediction accuracy for the cases of primary interest in this study, namely medium-sized and large fires. Providing this information to decision-makers responsible for firefighting operations could be valuable, facilitating the appropriate and timely allocation of resources in managing fire crises.

The problem addressed in this article is by no means exclusive to wildfire science, as it also affects many other sectors that rely on predictive models built from imbalanced datasets and used to guide managerial decisions. Examples include estimating the electricity generated by wind turbines, which is relevant for integrating wind energy into the power grid [28], forecasting customer demand [29], and other applications. Thus, a central purpose of this work is to contribute to the development of solutions that can assist decision-makers in such tasks, highlighting their relevance to domains far broader than wildfire science.

This article is organized as follows. Section 2 reviews related work in the domains of burned area prediction and imbalanced regression. Section 3 describes the study area and the Montesinho dataset. Section 4 details the preprocessing techniques, the machine learning algorithms employed, and the overall research methodology. Section 5 presents and discusses the results. Finally, Section 6 concludes the article.

2. Related work

2.1. Burned area prediction

In 2007, Cortez and Morais [10] applied machine learning algorithms to predict the final burned area of wildfires at the time of ignition in Montesinho Natural Park, Portugal. The dataset, collected between 2000 and 2003, primarily comprises meteorological variables, along with four additional features related to time and location. The authors evaluated several models, with the best performance achieved by the Support Vector Machines (SVM) algorithm. However, the model proved effective mainly in predicting the burned area of small fires. This study has inspired numerous researchers in the years since its publication, many of whom have sought to develop models with improved accuracy.

Castelli et al. [30] applied Geometric Semantic Genetic Programming (GS-GP) to the same dataset. This variant of genetic programming differs from standard GP by employing geometric semantic operators in place of the traditional crossover and mutation operators. Their implementation of GS-GP [31,32] was tested and compared against standard GP and other machine learning methods, yielding significantly improved results in terms of lower mean absolute error between the predicted and actual values.

As some studies [10,30] have reported poor performance of neural network models on the Montesinho wildfires dataset, Storer and Green [33] attempted to optimize the neural network architecture through extensive experimentation with various combinations of activation functions and hidden nodes. Additionally, Particle Swarm Optimization was employed to determine the network weights, replacing the conventional use of backpropagation. The authors report a significant improvement in model performance, both in comparison to a neural network trained using backpropagation and to the use of Particle Swarm Optimization alone [34].

Xie and Peng [35] examined the problem from both regression and classification perspectives. Various machine learning algorithms were evaluated, with Random Forest demonstrating the best performance in predicting burned area, and Extreme Gradient Boosting excelling in classification. Both algorithms are ensemble learning methods, meaning they combine the predictions of multiple base learners to enhance overall performance.

Yazici and Taskin [15] conducted a comprehensive review of studies focused on predicting the area burned in wildfires and contributed to the field by applying Bayesian optimization to identify optimal hyperparameters across several machine learning methods. This approach led to improvements in both prediction accuracy and training efficiency.

Other researchers have also utilized the Montesinho dataset, although with objectives other than predicting the burned area. Li et al. [36] investigated the locations most susceptible to wildfires and identified the most important factors influencing the extent of the burned area. Dong et al. [37] employed clustering and classification techniques to develop a model focused on the spatial and temporal characteristics of wildfires in the Montesinho region.

Li et al. [38] approached the problem from a time series perspective, proposing a Long Short-Term Memory (LSTM) model enhanced with an Attention mechanism to assign greater importance to the most relevant information. The authors reported that their model achieved high predictive accuracy and outperformed several baseline models.

Wood [39] obtained results comparable to those of traditional machine learning models using a transparent, open-box approach. However, the author contends that this model offers a key advantage: it enables the identification of specific combinations of wildfire records that contribute to individual predictions.

Despite their substantial contributions to the topic of burned-area prediction, existing studies commonly overlook large wildfires, which, although infrequent, account for an extensive share of the total burned area. Furthermore, in most studies models are assessed using only standard evaluation metrics, which represents a limitation because it provides little insight into how they perform specifically for large, high-impact wildfires. These two factors justify revisiting the Montesinho Park dataset with a particular focus on improving predictions for such fires.

2.2. Imbalanced regression

Imbalanced learning addresses datasets with an unequal distribution of data. In classification tasks, this refers to scenarios where one or more classes are significantly underrepresented [9]. In the context of regression problems, the term imbalanced regression is used to describe regression tasks where certain ranges of the continuous target variable are sparsely represented [9].

Different methods can be employed to address the problem of imbalanced learning, including algorithm-level methods, data-level methods, and cost-sensitive learning [40]. Among these, data-level methods stand out for their versatility. They involve modifying the distribution of the training data through oversampling or undersampling before applying a machine learning algorithm. As a preprocessing step, data-level techniques are independent of the specific algorithm used, allowing for the subsequent application of a wide range of popular algorithms, such as Neural Networks, Random Forest, or Gradient Boosting. For instance, the review studies by Jain et al. [4] and Ejaz and Choudhury [41] analyzed the wildfire science literature on machine learning applications, finding that these algorithms are among the most commonly used for wildfire predictions, independently of data-level preprocessing methods.

Furthermore, data-level methods have demonstrated success across a wide range of research domains [42], such as: information systems, computer science, biomedical engineering, multidisciplinary sciences, operations research management science, remote sensing, and environmental sciences, among others.

The considerations outlined above motivated our decision to employ data-level methods to tackle the problem at hand. Specifically, we employed Geometric SMOTE for regression [43] and WSMOTER [44], based on the characteristics of our problem. Both methods are oversampling techniques that augment the training set through the generation of synthetic samples. They were designed for regression tasks to enhance the representation of valuable samples which in our case are the larger wildfires. A detailed description of these oversamplers is provided in Section 4. Additionally, we also applied random undersampling to remove less informative samples, thereby improving the model's focus on the most relevant data.

3. Study area and data

The Montesinho Natural Park spans an area of 74,224.89 hectares and is located in northeastern Portugal, near the border with Spain. Although the risk of fire ignition in the park is not among the highest in the country [45], it is situated in a region where the likelihood of a resulting burned area being medium-sized or larger is significant due to characteristics such as population density, land cover, and distance to roads [46].

A detailed description of the Montesinho Natural Park, including maps and information on its location, landscape, climate, and topography, can be found in the studies by Cortez and Morais [10] and Castelli et al. [30].

Cortez and Morais [10] utilized a dataset collected between 2000 and 2003. Subsequent research has employed the same dataset in efforts to enhance machine-learning model accuracy. However, given the time that has passed since the original study, there is a clear need to update the dataset to incorporate more recent data.

Our study was conducted using data on wildfires that occurred between 2001 and 2022. The data were obtained from the Portuguese Forest Fire Information Management System [47].

As a preliminary observation, we highlight the distribution of burned area data, as shown in Fig. 1. The dataset exhibits a pronounced class imbalance, with only 59 medium and large fires recorded. This absolute number underscores the scarcity of such cases.

Regarding the features, we selected variables with both spatial and temporal characteristics: Latitude, Longitude, X, Y, Municipality, Month, and Hour. The X and Y features represent the coordinates of the fire ignition point in the PT-TM06/ETRS89 coordinate system. The Municipality variable indicates one of the two municipalities in which the Montesinho Park is located: Vinhais or Bragança. Additionally, we included the variable RNMPF, which specifies whether the fire originated within one of the forest perimeters of Montesinho Park.

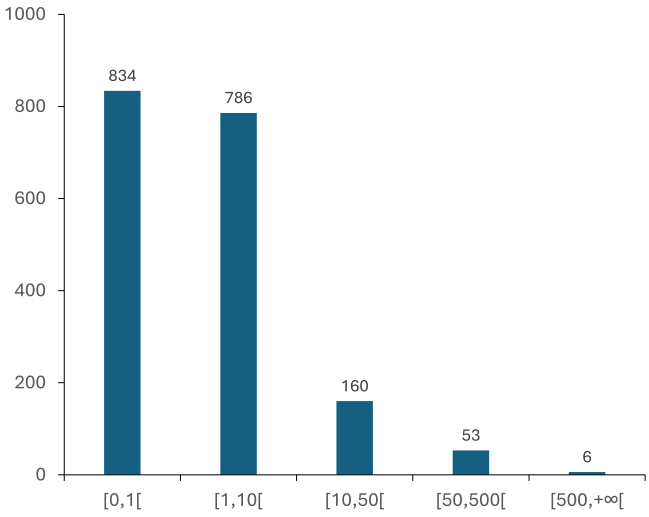

Fig. 1. Number of wildfires by burned area category.

Finally, we included seven features from Canada's widely used Fire Weather Index (FWI) System: DSR (Daily Severity Rating), FWI (Fire Weather Index), ISI (Initial Spread Index), DC (Drought Code), DMC (Duff Moisture Code), FFMC (Fine Fuel Moisture Code), and BUI (Buildup Index).

The data available in the SGIF dataset also enabled us to add four additional features to our dataset after appropriate preprocessing: First Intervention, Number of Fires in Municipality, Number of Fires in NUT III, and Number of Fires in NUT II. The First Intervention feature represents the time elapsed between fire detection and the initiation of firefighting efforts. The other three features capture the number of active wildfires occurring simultaneously at different administrative levels of increasing scale: municipality, NUT III region, and NUT II region. NUT refers to the Nomenclature of Territorial Units for Statistics used by the European Union. In the case of Montesinho Park, it is part of the Trás-os-Montes region (NUT III) and the Norte region (NUT II). These features aim to comprehensively reflect the operational context of firefighting at the time of ignition. From a theoretical standpoint, the simultaneous occurrence of multiple wildfires in the same region can hinder suppression efforts, as material and human resources are inherently limited.

Furthermore, we enriched our dataset with additional features from other sources. For example, we incorporated the Palmer Drought Severity Index (PDSI), obtained from the Portuguese Institute for Sea and Atmosphere (IPMA – Instituto Português do Mar e da Atmosfera, 2023), which enables the identification of drought periods.

We also obtained population density data for the municipalities of Vinhais and Bragança from Statistics Portugal [48]. The data were grouped in five-year intervals. Accordingly, we used population data from 2001 for all fires that occurred between 2001 and 2006, data from 2006 for fires between 2006 and 2011, and so on.

Additionally, we integrated several meteorological variables from the Visual Crossing Weather service [49] into our dataset: maximum temperature, minimum temperature, humidity, and precipitation (on the day of the fire alert). Furthermore, wind gust and wind speed (at the time of the fire alert) were also included. Finally, we added wind gust mean, wind speed mean, and wind rotation (average wind rotation per hour), which were calculated from hourly data over a 24-hour period.

Table 1
Features of the Montesinho Natural Park wildfires dataset.

Feature	Meaning
Latitude	Latitude of the ignition point
Longitude	Longitude of the ignition point
X	X of the ignition point in PT-TM06/ETRS89 coordinate system
Y	Y of the ignition point in PT-TM06/ETRS89 coordinate system
DSR	Daily Severity Rating
FWI	Fire Weather Index
ISI	Initial Spread Index
DC	Drought Code
DMC	Duff Moisture Code
FFMC	Fine Fuel Moisture Code
BUI	Buildup Index
First intervention	Time to first suppression intervention since ignition detection
RNMPF	Ignition point located in the National Network of National Forests
Population density	Population density of the municipality
PDSI	Palmer Drought Severity Index
Number fires municipality	Number of active fires in the municipality
Number fires NUT III	Number of active fires in the region NUT III
Number fires NUT II	Number of active fires in the region NUT II
TempMax	Maximum temperature
TempMin	Minimum temperature
Humidity	Relative humidity
Precip	Precipitation
Windgust	Wind gust at the time of fire alert
Windspeed	Wind speed at the time of fire alert
Windgust mean	Wind gust mean (in a 24 hour period)
Windspeed mean	Wind speed mean (in a 24 hour period)
Wind rotation	Average wind rotation per hour (in a 24 hour period)
Month	Month
Hour	Hour
Municipality	Municipality of the ignition point

The dataset we constructed is similar to that used in Cortez and Morais [10], as both mainly rely on meteorological features. However, our dataset differs in two key aspects: it includes a larger number of wildfire records due to the extended time range encompassing more recent fire events, and it incorporates a greater number of features.

Priority was given to including meteorological features in our study for several reasons. First, previous research has demonstrated that burned area can be predicted using meteorological data [10,12,13]. Second, Mitsopoulos and Mallinis [1] analyzed a comprehensive set of variables across different categories influencing the occurrence of large fires: fire suppression and behavior, weather, topography, and landscape, and concluded that variables related to fire suppression and weather are the primary drivers explaining large fires. Third, our use of fire suppression data was limited due to the lack of information on fire causes, fire type, and suppression methods. Moreover, even if such data were available, it remains debatable whether it is appropriate to include variables such as the resources allocated to suppression efforts, since the goal of this work is to enhance early-stage prediction of burned area based solely on information available at the time of fire ignition.

Table 1 provides a summary of all the features used in this study.

4. Algorithms and methodology

4.1. Imbalanced regression algorithms

Three data-level methods suitable for regression tasks were employed in this study: Geometric SMOTE, WSMOTER, and Random Undersampling. The latter was combined with each of the oversampling techniques to address the large number of small fires present in the dataset. An introduction to each method is provided below. The source code for the imbalanced regression methods, along with the dataset, is available at https://drive.google.com/drive/folders/1NhAuRV9b0WRxQmUfp06i0lmPNmNPgOh9?usp=sharing.

4.1.1. Geometric SMOTE

Geometric SMOTE [50] is an enhancement of the original SMOTE (Synthetic Minority Over-sampling Technique) [51] that enables the generation of higher-quality synthetic data. Practically, this improvement is achieved by increasing diversity through the creation of synthetic instances within a broader region around two existing samples, rather than solely along the line segment connecting them. Fig. 2 illustrates this concept. In the figure, minority samples are shown as unfilled circles and majority samples as filled circles. A represents a minority sample, and B one of its nearest neighbors. SMOTE generates synthetic instances along the line segment connecting A and B, whereas Geometric SMOTE generates instances throughout the entire grey region. In practice, SMOTE can be seen as a particular case of Geometric SMOTE.

The final shape of the region where synthetic instances are generated is determined by two transformations applied to a hyper-sphere: deformation and truncation. Fig. 3 illustrates how these transformations define the shape. In the figure, both transformations are illustrated separately as regions A and B for clarity, to highlight the effects of each, although they occur simultaneously within the same generation region. These transformations are controlled by two hyperparameters of the algorithm, while a third hyperparameter determines the neighbor selection strategy. Optimal values for these three hyperparameters must be identified through tuning.

Once determined, the final shape defines a safe zone in the input space within which synthetic instances are generated. This modification to the generation mechanism is particularly important, as it enables Geometric SMOTE to overcome inefficiencies of the original SMOTE, including the production of noisy instances due to the selection of initial samples or their neighbors, and the generation of synthetic instances that closely resemble the originals, potentially increasing the risk of overfitting.

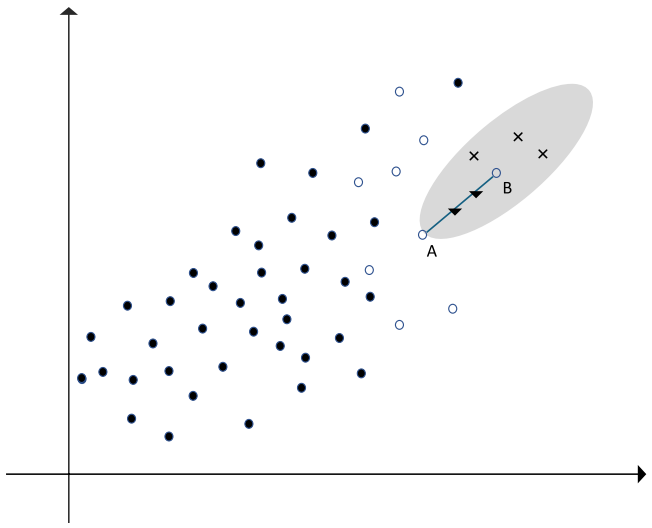

Fig. 2. Comparison of synthetic sample generation in SMOTE and Geometric SMOTE.

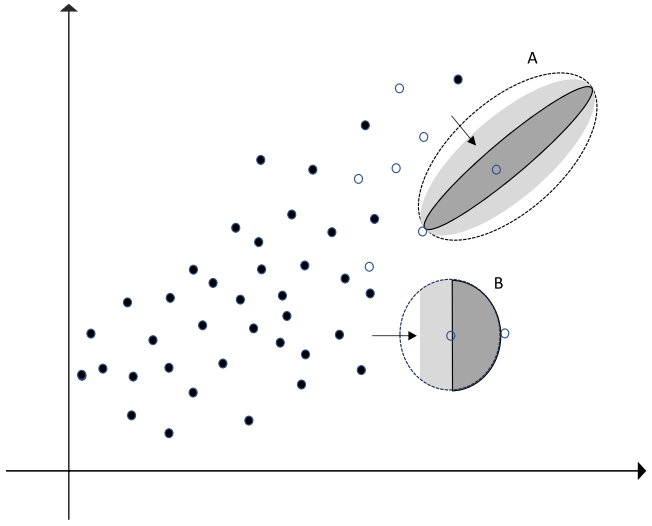

Fig. 3. Geometric SMOTE: region shaped by deformation (A) and truncation (B).

Although Geometric SMOTE was originally developed for classification tasks, it was subsequently adapted for regression problems [43], and this version is the one employed in our study.

4.1.2. WSMOTER

We also used WSMOTER [44] as an alternative oversampler to Geometric SMOTE. Although it is also based on SMOTE, WSMOTER possesses distinct characteristics that differentiate it from both the original SMOTE and Geometric SMOTE. Its main innovation lies in the method used to select samples as generators of synthetic instances, making it well-suited for regression tasks.

The sample selection strategy is performed without relying on predefined classes or partitions. The procedure is as follows: First, each sample is assigned a weight based on its rarity, with samples from low-density regions receiving higher weights than those from denser regions. Second, a sample is selected according to these weights, so that samples with greater weight have a higher chance of being chosen. Third, a second sample is selected from the neighborhood of the first, considering both feature space and target value. Finally, a synthetic sample is generated from the two selected real samples and added to the training set.

In the task of predicting burned area, historical wildfire records exhibit a large number of small fires and progressively fewer fires as the burned area increases. WSMOTER assigns a weight to each fire based on its rarity, giving higher weights to larger fires. Consequently, records of larger fires have a higher probability of being selected for interpolation with other samples, generating new synthetic instances that help improve model training.

4.1.3. Random undersampling

Undersampling methods perform the opposite function of oversamplers by removing less informative examples from the training set. In our task, these are the small fires, and undersampling can be applied randomly, provided it is restricted to this category. Given the high imbalance between large and small fires, we employed Random Undersampling alongside Geometric SMOTE and WSMOTER to achieve a better balance in the training dataset.

4.2. Machine learning algorithms

After resampling the training set to increase the representation of large wildfires and reduce the influence of the large number of small fires, the model was trained using the following machine learning algorithms: Decision Trees (DT), Random Forest (RF), Support Vector Regression (SVR), and Gradient Boosting (GB).

These algorithms were selected because they are among the most widely used and have consistently demonstrated strong performance across various domains in wildfire science [4]. Nonetheless, other machine learning algorithms could have also been employed, as the resampling techniques used in this study are not specific to the algorithms applied.

The machine learning algorithms were implemented using the versions available in the Scikit-learn [52] and XGBoost [53] libraries, as all programming was conducted in the Python language. It is worth noting that XGBoost was used exclusively for the Gradient Boosting algorithm. Detailed descriptions of the algorithms can be found in the official documentation of the respective libraries.

All experiments were run on a computer with an AMD Ryzen 7 4800H processor with Radeon Graphics (2.90 GHz), an NVIDIA GeForce GTX 1650 Ti (4 GB) graphics card, and 16 GB of RAM. The operating system was Windows 10.

4.3. Methodology

The experimental phase of our work was carried out using 5-fold stratified cross-validation. The procedure can be summarized as follows: (1) the dataset is split into 5-fold, and (2) in each iteration of the training/testing process, four folds are used for training while the remaining fold is used for testing. This process ensures that each fold is used exactly once as a test set.

The stratified version of cross-validation was adopted to address the high imbalance ratio in the data. Stratification ensures that the proportion of large, medium, and small wildfires is consistently maintained across all folds. This is crucial for having representative training and test sets, which is particularly important in imbalanced scenarios.

Various combinations of hyperparameter values were tested for each machine learning algorithm in an effort to optimize model performance. Table 2 summarizes the hyperparameter configurations considered.

Regarding imbalanced regression techniques, different oversampling ratios were evaluated for both WSMOTER and Geometric SMOTE. The ratios considered were drawn from the set $\{2, 3, 5\}$. However, different oversampling strategies were adopted for each technique. For Geometric SMOTE, synthetic instances were generated exclusively from

Table 2
Hyperparameters of the Machine Learning algorithms.

	Hyperparameters	Values
Decision Trees	min samples split	{2,5,10}
	max features	{1.0,0.75,0.5}
Random Forest	number of trees	{500}
	min samples split	{2,5,10}
	max features	{1.0,0.75,0.5}
Support Vector Regression	kernel	{'rbf','poly'}
	C	{1,10,100,300}
Gradient Boosting	number of trees	{500}
	learning rate	{0.1,0.3,0.6}
	max depth	{2,6,10}

Table 3
Hyperparameters of the Geometric SMOTE.

	Values
Deformation	{0. 0.25,0.5,0.75,1.}
Truncation	{−1.,−0.75,−0.5,−0.25,0.,0.25,0.5,0.75,1.}
Neighbor selection strategy	{'minority','majority','combined'}

wildfires with a burned area greater than 50 hectares. In contrast, WSMOTER operated over a broader target domain, including all wildfires with a burned area above the median of the target values. Additionally, a grid search was conducted to tune the three hyperparameters of Geometric SMOTE in order to identify the optimal configuration. The values explored for each hyperparameter are summarized in Table 3.

The resampling of the training sets was complemented by the application of Random Undersampling to reduce the number of records corresponding to smaller fires. The undersampling targeted samples with burned area values below the threshold established for oversampling. In the case of WSMOTER, the number of these samples was reduced by 50%. For Geometric SMOTE, the reduction was calibrated to ensure that wildfires with a burned area greater than 50 hectares accounted for 20% of the training set.

The best results from all possible combinations are presented in the tables of results in Section 5.

5. Results and discussion

Traditionally, the performance of regression models is evaluated using standard metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). A new model is generally considered successful if it achieves lower values on either of these metrics. However, it is important to clarify how errors are distributed across the full range of predictions, from the smallest to the largest wildfires. Relying solely on global MAE and RMSE values is insufficient for determining whether one model is truly better than another. For example, consider a hypothetical model that predicts a constant burned area of 1 hectare for every wildfire. Such a model would yield an MAE of 9.2 which might seem quite impressive because this MAE value is even lower than the best results reported in prior studies. Nevertheless, this model would be entirely ineffective at predicting large burned areas, making it practically useless for identifying large wildfire events.

Based on the considerations discussed above, the results are presented according to wildfire size. To this end, three wildfire size classes were defined: small, medium, and large. As there is no universally accepted threshold in the literature for categorizing a wildfire as "large" we adopted the classification criteria proposed by Mitsopoulos and Mallinis [1]. According to this criterion, small wildfires are those with a burned area of less than 50 hectares, medium wildfires have burned areas between 50 and 500 hectares, and large wildfires are those exceeding 500 hectares.

Although this decision complicates comparisons with previous studies, it is essential for obtaining an accurate understanding of prediction errors across the entire domain and in particular for large wildfires.

Table 4
Mean absolute error using Random Forest.

Burned area	None	WSMOTER	Geometric SMOTE
Small (<50)	10.3	24.7	23.9
Medium (⩾50 and <500)	102.1	83.0	90.9
Large (⩾500)	753.7	718.5	613.9

Table 5
Mean absolute error using Decision Trees.

Burned area	None	WSMOTER	Geometric SMOTE
Small (<50)	12.3	22.2	22.4
Medium (⩾50 and <500)	109.7	100.5	120.4
Large (⩾500)	643.7	624.9	457.3

Table 6
Mean absolute error using Support Vector Regression.

Burned area	None	WSMOTER	Geometric SMOTE
Small (<50)	3.6	24.7	23.8
Medium (⩾50 and <500)	119.5	84.4	100.3
Large (⩾500)	808.6	769.7	754.1

Table 7
Mean absolute error using Gradient Boosting.

Burned area	None	WSMOTER	Geometric SMOTE
Small (<50)	10.8	20.1	17.9
Medium (⩾50 and <500)	103.7	86.2	96.1
Large (⩾500)	788.4	720.3	606

Before presenting the results, it is important to clarify what is meant by a model being "better" in a context where accurate predictions for medium-sized and large fires are especially desirable. A fundamental challenge lies in the trade-off that often exists between prediction accuracy for large and small fires: improvements in predicting large fires frequently come at the cost of reduced performance on smaller ones. To address this, we defined a threshold for what constitutes an acceptable error in small fires, which was arbitrarily set at 25 hectares. Reported results, therefore, represent the best performance for large wildfires without exceeding this predefined error limit for small fires. In the specific case of the WSMOTER method, we chose to report the best results obtained for medium-sized wildfires, as this resampling technique tends to perform particularly well in this category. This advantage may be attributed to the broader range available for generating synthetic instances within the medium fire class. Interestingly, for two of the four algorithms evaluated, the hyperparameters that yielded the best results for medium-sized fires were also optimal for large wildfires.

Tables 4–7 summarize the best results achieved using the resampling techniques and learning algorithms employed in this study. In the tables, 'None' is used as a label to indicate that no resampling was performed.

An initial analysis of the tables confirms an improvement in model performance for medium and large wildfires. This enhancement, however, is accompanied by a decrease in prediction accuracy for smaller fires. Despite this drawback, the reduction in accuracy for small fires should not be a major concern, as it is largely offset by the improved ability to anticipate larger wildfires. Although large wildfires represent only 3.2% of the total number of fires in the dataset, they account for 65.8% of the total burned area. As such, improving the prediction of large wildfire occurrences has the potential to significantly reduce the total burned area, provided it is accompanied by effective fire suppression strategies.

Surprisingly, although not without precedent, the best result for large wildfires, both in absolute and percentage terms, was achieved using a simple Decision Tree. A similar outcome was observed, for example, in Coffield et al. [12]. However, this result was obtained at

Table 8
Wilcoxon Signed-Rank Test: test statistic, p-value and significance.

	Statistic	p-value	Significance
DT + WSMOTER	541	0.00941753	True
DT + Geometric SMOTE	725	0.31231498	False
RF + WSMOTER	173	7.69409874e−08	True
RF + Geometric SMOTE	258	2.21682864e−06	True
SVR + WSMOTER	4	2.93557693e−11	True
SVR + Geometric SMOTE	201	2.43262794e−07	True
GB + WSMOTER	388	0.00017589	True
GB + Geometric SMOTE	593	0.02752371	True

the expense of poor performance on medium-sized wildfires, making it a unique case among the algorithms evaluated. In comparison, algorithms such as Random Forest and Gradient Boosting provided more balanced performance. These approaches delivered strong results for large wildfires while also improving predictions for medium-sized fires, with error reductions ranging from 7.3% to 18.7%. For large wildfires in particular, the best results were generally achieved when the models were combined with Geometric SMOTE resampling, yielding improvements between 18.5% and 23.1%.

With regard to Support Vector Regression, the results tables clearly indicate that, without resampling, this algorithm performs best on small fires and worst on large ones, with MAE values of 3.6 and 808.6, respectively. This helps explain why Support Vector Machines achieved the best results in the study by Cortez and Morais [10]. The low error for small fires, combined with the high proportion of such fires in the dataset, resulted in the lowest overall MAE across all algorithms compared. However, this performance is largely due to the model's systematic prediction of low burned area values for all wildfires. While this leads to impressive MAE figures, it has limited practical value because the model fails to accurately predict large fires. Although the use of resampling techniques does not completely solve this issue with this algorithm, it does help improve performance, particularly for medium-sized wildfires. In fact, error reductions of 29.4% and 16.1% were observed in the two resampling approaches applied, when compared to not using resampling at all.

Finally, Table 7 presents the results for Gradient Boosting, where Geometric SMOTE stands out by achieving a 23.1% reduction in error for large wildfires, with a smaller negative impact on the performance for small wildfires compared to the other algorithms.

5.1. Statistical tests

The non-parametric Wilcoxon Signed-Rank Test was applied for statistical validation of the results, as it does not require the assumption of normally distributed data.

The null hypothesis tested with the Wilcoxon Signed-Rank Test states that the prediction errors for medium and large wildfires produced by the integrated resampling–machine learning model are not significantly different from those of the baseline model without resampling. This focus on medium and large wildfires reflects the study's emphasis on accurately predicting higher-magnitude wildfire events.

Because the study employed two resampling methods and four machine learning algorithms, a total of eight resampling–algorithm combinations were evaluated. This configuration resulted in eight pairwise comparisons between each combination and the baseline model.

Table 8 reports the Wilcoxon Signed-Rank test statistic and the corresponding p-value for the eight combinations formed by pairing each machine-learning algorithm with either WSMOTER or Geometric SMOTE. The decisions regarding the rejection of the null hypothesis for each comparison are also summarized in the table.

The null hypothesis was rejected at the significance level of $\alpha = 0.05$ for all but one of the resampling–algorithm combinations. Therefore, the observed differences in predictive performance are statistically significant for seven of the eight combinations.

All computations for the Wilcoxon Signed-Rank Test were performed using the SciPy library [54].

5.2. Limitations

Our framework was applied to the Montesinho Park, which is a geographically limited area. Although it is expected that applying the framework to other regions would contribute to improving predictions of medium and large fires, the observed improvements may not fully transfer to different settings. This limitation arises from contextual differences such as fuel types, climate, or management practices. In such situations, the inclusion of additional features that better describe the local environment may enhance model performance and deserves further investigation.

Regarding the oversamplers applied, Geometric SMOTE enhances the diversity of the generated synthetic instances, improving the representation of rare patterns. However, it requires hyperparameter tuning to achieve optimal results, which can be computationally demanding. WSMOTER, in contrast, does not require hyperparameter tuning and performs well when rare cases are widely dispersed, but it is less effective at increasing the diversity of synthetic instances compared to Geometric SMOTE.

6. Conclusion

The research community has been using machine learning algorithms to predict the final size of a fire. However, results are often under expectations, particularly with regard to large wildfires. Although these events are relatively rare, they account for a substantial portion of the total burned area and have significant impacts on both human populations and the environment. Moreover, in the context of climate change, it is reasonable to anticipate a rise in the frequency of large-scale fires in the future. Therefore, improving the ability to predict the occurrence of medium-sized and large wildfires is highly desirable, as it could be a valuable tool in efforts to suppress and mitigate their consequences.

Our approach utilizes recent resampling techniques to generate synthetic instances representing the most relevant cases, thereby improving the algorithms' capacity to learn from them. Depending on the algorithm used, this strategy can reduce prediction error for large fires by up to 29%. More balanced solutions, which perform well across large, medium, and small fire categories, achieve error reductions of approximately 20% for large fires. This improvement in predicting large fires is typically also accompanied by a reduction in error for medium-sized fires, though the extent of this reduction varies depending on the algorithm and the user's prioritization of accuracy for large versus medium fires. Achieving optimal performance in both categories simultaneously often proves challenging. While the approach results in less accurate predictions for small fires, this drawback is largely outweighed by the enhanced ability to anticipate large fires, which is crucial for enabling earlier and more effective suppression efforts.

Furthermore, any research in this field must take into account the importance and inherent challenges of predicting rare events, such as large wildfires. This consideration also extends to other domains of interest, where approaches similar to the one adopted in this study may prove valuable, such as predicting fine particle concentrations to assess human exposure to smoke. More broadly, applying this approach could

be beneficial in other contexts where accurate prediction of events from limited data is essential.

Finally, it is important to emphasize that our work focuses on enhancing model performance in the most challenging cases, whereas other studies have aimed to improve models from a more general perspective. As future work, combining both approaches may be a promising direction for developing robust and effective predictive models.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project - UID/04152/2025 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS - https://doi.org/10.54499/UID/04152/2025 (2025-01-01/2028-12-31) and UID/PRR/04152/2025 https://doi.org/10.54499/UID/PRR/04152/2025 (2025-01-01/ 2026-06-30). This work was funded by research contract CEECIND/01378/2017/CP1430/CT0005 and by FEDER funds through the Operational Competitiveness Factors Program - COMPETE and by national funds through FCT – Portuguese Foundation for Science and Technology within the scope of the project UIDB/50027/2025.

Data availability

The source code for the imbalanced regression methods, together with the dataset used in this study, is publicly available at: https://drive.google.com/drive/folders/1NhAuRV9b0WRxQmUfp06i0lmPNmNPgOh9?usp=sharing.

References

[1] I. Mitsopoulos, G. Mallinis, A data-driven approach to assess large fire size generation in Greece, Nat. Hazards 88 (2017) 1591–1607, http://dx.doi.org/10.1007/s11069-017-2934-z.

[2] F.H. Johnston, S.B. Henderson, Y. Chen, J.T. Randerson, M. Marlier, R.S. DeFries, P. Kinney, D.M. Bowman, M. Brauer, Estimated global mortality attributable to smoke from landscape fires, Environ. Health Perpectives 120 (2012) 695–701, http://dx.doi.org/10.1289/ehp.1104422.

[3] S. Meier, R.J. Elliott, E. Strobl, The regional economic impact of wildfires: Evidence from southern Europe, J. Environ. Econ. Manag. 118 (2023) 102787, http://dx.doi.org/10.1016/j.jeem.2023.102787.

[4] P. Jain, S.C. Coogan, S.G. Subramanian, M. Crowley, S. Taylor, M.D. Flannigan, A review of machine learning applications in wildfire science and management, Environ. Rev. 28 (4) (2020) 478–505, http://dx.doi.org/10.1139/er-2020-0019.

[5] I. Silva, M. Valle, L. Barros, J. Meyer, A wildfire warning system applied to the state of acre in the Brazilian amazon, Appl. Soft Comput. 89 (2020) 106075, http://dx.doi.org/10.1016/j.asoc.2020.106075.

[6] Y. Kang, E. Jang, J. Im, C. Kwon, A deep learning model using geostationary satellite data for forest fire detection with reduced detection latency, GIScience Remote. Sens. 59 (1) (2022) 2019–2035, http://dx.doi.org/10.1080/15481603.2022.2143872.

[7] K. Henderson, R.K. Chakrabortty, A machine learning predictive model for bushfire ignition and severity: The study of Australian black summer bushfires, Decis. Anal. J. 14 (2025) 100529, http://dx.doi.org/10.1016/j.dajour.2024.100529.

[8] N. Paul, J. Yao, K.E. McLean, D.M. Stieb, S.B. Henderson, The Canadian optimized statistical smoke exposure model (CanOSSEM): A machine learning approach to estimate national daily fine particulate matter (PM2.5) exposure, Sci. Total Environ. 850 (2022) 157956, http://dx.doi.org/10.1016/j.scitotenv.2022.157956.

[9] A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk, F. Herrera, Cost-sensitive learning, in: Learning from Imbalanced Data Sets, Springer International Publishing, Cham, 2018, pp. 63–78, http://dx.doi.org/10.1007/978-3-319-98074-4.

[10] P. Cortez, A.d.R. Morais, A data mining approach to predict forest fires using meteorological data, in: J.M. Neves, M.F. Santos, J.M. Machado (Eds.), New Trends in Artificial Intelligence : Proceedings of the 13th Portuguese Conference on Artificial Intelligence, EPIA 2007, Guimarães, Portugal, 2007, pp. 512–523, URL https://hdl.handle.net/1822/8039.

[11] F. Li, Q. Zhu, W.J. Riley, L. Zhao, L. Xu, K. Yuan, M. Chen, H. Wu, Z. Gui, J. Gong, J.T. Randerson, AttentionFire_v1.0: interpretable machine learning fire model for burned-area predictions over tropics, Geosci. Model. Dev. 16 (3) (2023) 869–884, http://dx.doi.org/10.5194/gmd-16-869-2023.

[12] S.R. Coffield, C.A. Graff, Y. Chen, P. Smyth, E. Foufoula-Georgiou, J.T. Randerson, Machine learning to predict final fire size at the time of ignition, Int. J. Wildland Fire 28 (2019) 861–873, http://dx.doi.org/10.1071/WF19023.

[13] H. Liang, M. Zhang, H. Wang, A neural network model for wildfire scale prediction using meteorological factors, IEEE Access 7 (2019) 176746–176755, http://dx.doi.org/10.1109/ACCESS.2019.2957837.

[14] A.M. Irawan, M. Vall-llossera, C. López-Martínez, A. Camps, D. Chaparro, G. Portal, M. Pablos, Burned area prediction in southern Asia using machine learning with land and atmospheric parameters, in: IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium, 2023, pp. 2930–2933, http://dx.doi.org/10.1109/IGARSS52108.2023.10283214.

[15] K. Yazici, A. Taskin, A comparative Bayesian optimization-based machine learning and artificial neural networks approach for burned area prediction in forest fires: an application in Turkey, Nat. Hazards 119 (2023) 1883–1912, http://dx.doi.org/10.1007/s11069-023-06187-4.

[16] Global Wildfire Information System, Global Monthly Burned Area, 2025, https://gwis.jrc.ec.europa.eu/apps/country.profile/downloads. (19 Accessed November 2025).

[17] L. Giglio, L. Boschetti, D.P. Roy, M.L. Humber, C.O. Justice, The collection 6 MODIS burned area mapping algorithm and product, Remote Sens. Environ. 217 (2018) 72–85, http://dx.doi.org/10.1016/j.rse.2018.08.005.

[18] E.J. Hanan, J. Ren, C.L. Tague, C.A. Kolden, J.T. Abatzoglou, R.R. Bart, M.C. Kennedy, M. Liu, J.C. Adam, How climate change and fire exclusion drive wildfire regimes at actionable scales, Environ. Res. Lett. 16 (2) (2021) 024051, http://dx.doi.org/10.1088/1748-9326/abd78e.

[19] A. Rovithakis, M.G. Grillakis, K.D. Seiradakis, C. Giannakopoulos, A. Karali, R. Field, M. Lazaridis, A. Voulgarakis, Future climate change impact on wildfire danger over the mediterranean: the case of Greece, Environ. Res. Lett. 17 (4) (2022) 045022, http://dx.doi.org/10.1088/1748-9326/ac5f94.

[20] T. van der Schriek, K.V. Varotsos, A. Karali, C. Giannakopoulos, Wildfire Burnt Area and associated greenhouse gas emissions under future climate change scenarios in the mediterranean: Developing a robust estimation approach, Fire 7 (9) (2024) http://dx.doi.org/10.3390/fire7090324.

[21] C.S. Gannon, N.C. Steinberg, A global assessment of wildfire potential under climate change utilizing keetch-byram drought index and land cover classifications, Environ. Res. Commun. 3 (3) (2021) 035002, http://dx.doi.org/10.1088/2515-7620/abd836.

[22] E. Brown, B.P. Hunt, Cumulative effects of fire in the fraser river basin on freshwater quality and implications for the salish sea, Sci. Total Environ. 978 (2025) 179416, http://dx.doi.org/10.1016/j.scitotenv.2025.179416.

[23] C.P. Brucker, B. Livneh, F.L. Rosario-Ortiz, F. Yao, A.P. Williams, W.C. Becker, S.K. Kampf, B. Rajagopalan, Wildfires drive multi-year water quality degradation over the western United States, Commun. Earth & Environ. 6 (489) (2025) http://dx.doi.org/10.1038/s43247-025-02427-6.

[24] K. Gajendiran, S. Kandasamy, M. Narayanan, Influences of wildfire on the forest ecosystem and climate change: A comprehensive study, Environ. Res. 240 (2024) 117537, http://dx.doi.org/10.1016/j.envres.2023.117537.

[25] F. Guerrero, L. Espinoza, V. Vidal, C. Carmona, P. Krecl, A.C. Targino, M.F. Ruggeri, M. Toledo, Black carbon and particulate matter concentrations amid central Chile's extreme wildfires, Sci. Total Environ. 951 (2024) 175541, http://dx.doi.org/10.1016/j.scitotenv.2024.175541.

[26] M. Filonchyk, M.P. Peterson, L. Zhang, L. Zhang, Y. He, Estimating air pollutant emissions from the 2024 wildfires in Canada and the impact on air quality, Gondwana Res. 140 (2025) 194–204, http://dx.doi.org/10.1016/j.gr.2024.12.012.

[27] E. Da Ponte, F. Alcasena, T. Bhagwat, Z. Hu, L. Eufemia, A.P. Dias Turetta, M. Bonatti, S. Sieber, P.-L. Barr, Assessing wildfire activity and forest loss in protected areas of the Amazon basin, Appl. Geogr. 157 (2023) 102970, http://dx.doi.org/10.1016/j.apgeog.2023.102970.

[28] H. Elmousalami, H.H. Elmesalami, M. Maxi, A.A.K.M. Farid, N. Elshaboury, A comprehensive evaluation of machine learning and deep learning algorithms for wind speed and power prediction, Decis. Anal. J. 13 (2024) 100527, http://dx.doi.org/10.1016/j.dajour.2024.100527.

[29] M. Karunanithi, P. Chatasawapreeda, T.A. Khan, A predictive analytics approach for forecasting bike rental demand, Decis. Anal. J. 11 (2024) 100482, http://dx.doi.org/10.1016/j.dajour.2024.100482.

[30] M. Castelli, L. Vanneschi, A. Popovič, Predicting Burned Areas of forest fires: an artificial intelligence approach, Fire Ecol. 11 (2015) 106–118, http://dx.doi.org/10.4996/fireecology.1101106.

[31] M. Castelli, S. Silva, L. Vanneschi, A C++ framework for geometric semantic genetic programming, Genet. Program. Evolvable Mach. 16 (2015) 73–81, http://dx.doi.org/10.1007/s10710-014-9218-0.

[32] L. Vanneschi, M. Castelli, L. Manzoni, S. Silva, A new implementation of geometric semantic GP and its application to problems in pharmacokinetics, in: K. Krawiec, A. Moraglio, T. Hu, A.Ş. Etaner-Uyar, B. Hu (Eds.), Genetic Programming, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 205–216.

[33] J. Storer, R. Green, PSO trained neural networks for predicting forest fire size: A comparison of implementation and performance, in: 2016 International Joint Conference on Neural Networks, IJCNN, 2016, pp. 676–683, http://dx.doi.org/10.1109/IJCNN.2016.7727265.

[34] Y. Wang, J. Wang, W. Du, C. Wang, Y. Liang, C. Zhou, L. Huang, Immune particle swarm optimization for support vector regression on forest fire prediction, in: W. Yu, H. He, N. Zhang (Eds.), Advances in Neural Networks – ISNN 2009, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 382–390.

[35] Y. Xie, M. Peng, Forest fire forecasting using ensemble learning approaches, Neural Comput. Appl. 31 (2019) 4541–4550, http://dx.doi.org/10.1007/s00521-018-3515-0.

[36] H. Li, X. Fei, C. He, Study on most important factor and most vulnerable location for a forest fire case using various machine learning techniques, in: 2018 Sixth International Conference on Advanced Cloud and Big Data, CBD, 2018, pp. 298–303, http://dx.doi.org/10.1109/CBD.2018.00060.

[37] H. Dong, H. Wu, P. Sun, Y. Ding, Wildfire prediction model based on spatial and temporal characteristics: A case study of a wildfire in Portugal's Montesinho natural park, Sustainability 14 (16) (2022) http://dx.doi.org/10.3390/su141610107.

[38] Z. Li, Y. Huang, X. Li, L. Xu, Wildland Fire Burned Areas prediction using long short-term memory neural network with attention mechanism, Fire Technol. 57 (2021) 1–23, http://dx.doi.org/10.1007/s10694-020-01028-3.

[39] D.A. Wood, Prediction and data mining of burned areas of forest fires: Optimized data matching and mining algorithm provides valuable insight, Artif. Intell. Agric. 5 (2021) 24–42, http://dx.doi.org/10.1016/j.aiia.2021.01.004.

[40] G. Douzas, F. Bacao, F. Last, Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE, Inform. Sci. 465 (2018) 1–20, http://dx.doi.org/10.1016/j.ins.2018.06.056.

[41] N. Ejaz, S. Choudhury, A comprehensive survey of the machine learning pipeline for wildfire risk prediction and assessment, Ecol. Inform. 90 (2025) 103325, http://dx.doi.org/10.1016/j.ecoinf.2025.103325.

[42] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, G. Bing, Learning from class-imbalanced data: Review of methods and applications, Expert Syst. Appl. 73 (2017) 220–239, http://dx.doi.org/10.1016/j.eswa.2016.12.035.

[43] L. Camacho, G. Douzas, F. Bacao, Geometric SMOTE for regression, Expert Syst. Appl. 193 (2022) 116387, http://dx.doi.org/10.1016/j.eswa.2021.116387.

[44] L. Camacho, F. Bacao, WSMOTER: a novel approach for imbalanced regression, Appl. Intell. 54 (2024) 8789–8799, http://dx.doi.org/10.1007/s10489-024-05608-6.

[45] F.X. Catry, F.C. Rego, F. Bação, F. Moreira, Modeling and mapping wildfire ignition risk in Portugal, Int. J. Wildland Fire 18 (2009) 921–931, http://dx.doi.org/10.1071/WF07123.

[46] F. Moreira, F.X. Catry, F. Rego, F. Bacao, Size-dependent pattern of wildfire ignitions in Portugal: when do ignitions turn into big fires? Landsc. Ecol. 25 (2010) 1405–1417, http://dx.doi.org/10.1007/s10980-010-9491-0.

[47] SGIF – Sistema de Gestão de Informação de Incêndios Florestais, Statistical information on rural fires, 2023, www.icnf.pt/florestas/gfr/gfrgestaoinformacao/estatisticas. (Accessed 10 May 2023).

[48] Instituto Nacional de Estatística, IP – Portugal, Anuários estatísticos regionais (2001,2006,2011,2016,2021), 2023, https://www.ine.pt/. (Accessed 25 September 2023).

[49] Visual Crossing Corporation, Visual crossing weather (2001–2022), 2025, https://www.visualcrossing.com/. (Accessed 23 January 2025).

[50] G. Douzas, F. Bacao, Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE, Inform. Sci. 501 (2019) 118–135, http://dx.doi.org/10.1016/j.ins.2019.06.007.

[51] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Int. Res. 16 (1) (2002) 321–357.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[53] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD 16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 785–794, http://dx.doi.org/10.1145/2939672.2939785.

[54] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental algorithms for scientific computing in python, Nature Methods 17 (2020) 261–272, http://dx.doi.org/10.1038/s41592-019-0686-2.