Journal of Computation Physics and Earth Science
ISSN: 2776-2521 (online)
Volume 1, Number 1, April 2021, Page 24-31
https://journal.physan.org/index.php/jocpes/index

Visualization of Fire Weather Index in Aek Godang Based on the Machine Learning Approach

Kerista Tarigan[1], Edison Kurniawan[2], Sri Wahyuni[2]
[1]Department of Physics, FMIPA, Univesitas Sumatera Utara, Medan, Indonesia
[2]Badan Meteorologi Klimatologi dan Geofisika, Indonesia

Article Info

Article history:

Received March 9, 2021
Revised March 20, 2021
Accepted March 26, 2021

Keywords:

Fire Weather Index
Machine Learning
SVM

A B S T R A C T

Forest fires are a major natural issue, making temperate and environmental harm whereas dangering human lives. The examined and study for timberland fire had been worn out Aek Godang, Northern Sumatera, Indonesia. There are 26 hotspots in 2017 near Aek Godang, North Sumatera, Indonesia. In this consider, we utilize an information mining approach to prepare and test the information of woodland fire and Fire Weather Index (FWI) from meteorological information. The point of this ponders to anticipate the burned range and distinguish the woodland fire in Aek Godang ranges, North Sumatera. The result of this considers shown the Fire battling and avoidance movement may be one reason for the watched need of relationship. The reality that this dataset exists demonstrates that there's as of now a few exertions going into fire avoidance.

Corresponding Author:

Kerista Tarigan,
Department of Physics, FMIPA, Universitas Sumatera Utara, Medan, Indonesia
Email: kerista@usu.ac.id

1. INTRODUCTION

Every year Northern Sumatra of Indonesia spends hundreds of thousands in order to deal with the wildfire breakout. This situation now not solely motives monetary damage however can additionally disrupt the ecological stability by way of destroying vegetation and plants and fauna [1]. Wildfire is additionally accountable for air pollution and changes in climatic circumstance over the period of time [2]. Over the decade forest fire has turn out to be a major problem as it has endangered the lives of species. regardless of the massive charges concerned in controlling these dead fires, they are additionally a essential problem in forest fires [3]. The forests on the border of Aek Godang areas, North Sumatra had been badly affected and would be impacted by means of different areas in North Sumatra. The primary trouble of this study, how to computation the hotspot in this location to predict the woodland fire. Firefighters are conscious of how forest fires can be unpredictable [4]. However, if this data is obtained through them as a warning about the breakout in a timely manner then this form of phenomenon can be anticipated, controlled mainly can be prevented. There are many typical applied sciences that deal with wildfire hazard analysis. In this study, based on the description above, we are aiming to remedy this trouble via a historical analysis of woodland and land furnace facts and the usage of weather data to predict the extent of fires that have occurred. Then we also explored information mining strategies to locate out and predict the depth of wooded area and land fires [5]. Fast detection is a key component for controlling such a phenomenon. In achieving this, alternative

options are needed. one of them is the use of nearby sensor-based automatic equipment furnished by using a number of meteorological stations [6]. causing meteorological conditions (such as temperature and wind) to affect wooded area and land fires, as well as knowing what a furnace index, such as the Fire Weather Index (FWI), makes use of this data. FWI is primarily based on the Index Spread Index (ISI) about the spread of furnace and wind speed, then the Buildup Index (BUI) to calculate the quantity of gasoline that reasons a fire. All of this is used as a measure for the well-known index of hearth hazard in woodland areas. In this work, we conducted statistics exploration with a information mining (DM) strategy so that we may want to predict the place of forest fires and burned land [7]. In this study we use Support Vector Machines (SVM)[9][10] and four distinct feature selection setups (using spatial, temporal, FWI components and weather attributes), by carrying out tests on the latest real-world data, data collected from the northern Sumatera. The satisfactory configuration end result is the use of the SVM method with 4 meteorological enter parameters (namely relative humidity, rain, temperature, and wind) and is capable to predict burnt areas from several widely wide-spread small fires. So, this know-how is very supportive and useful in enhancing preventive motion and administration of firefighting sources (equipment and people).

.

2. DATA AND METHOD

The dataset of this study had been collect in BMKG of Aek Godang Station, North Sumatera from 2017 years[1], form the LAPAN based on the Satellite of NOAA[2] and from PKHL Direktorat Pengendalian Kebakaran Hutan[3]. There is more than 26 hotspot in the 2017 close to Aekgodang, Northern Sumatera was recorded[4].

The FWI file is an pointer measuring chimney escalated and combining the two going before components. In spite of the fact that the scale utilized is unmistakable for each issue of the FWI, the culminate fetched may too show additional extreme combustion conditions. At that point the distinctive imperative component is that the gasoline mugginess code requires memory (time slack) of the going before climate conditions: is 12 days for DMC, sixteen hours for FFMC, and 52 days for DC. Usually an basic pointer in figuring out the profundity of lush region and arrive fires that take place.

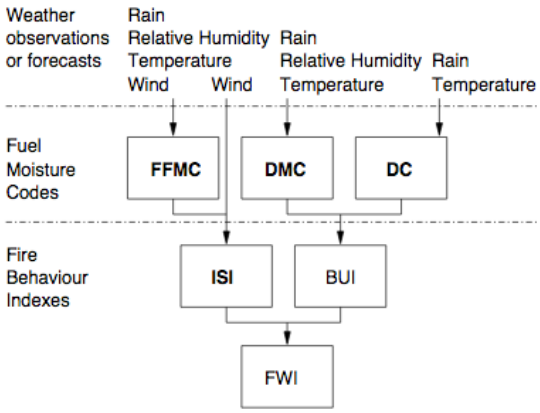


Figure 1. The Fire Weather Index structure [5]

A regression dataset D is made up of $k \in \{1, ..., N\}$ examples, each mapping an input vector $(x_1^k, \ldots . x_A^k)$ to a given target y_k. The error is given by : $e_k = y_k - \hat{y}_k$, where $\hat{y}_k$ represents the predicted value for the k input pattern. The overall performance is computed by a global metric, namely the *Mean Absolute Deviation* (MAD) and *Root Mean Squared* (RMSE)[6], which can be computed as eq.1.

$$MAD = 1/N \; x \; \sum_{i=1}^{N} | yk \; - \; \hat{y}k| \qquad RMSE = \sqrt{\sum_{i=1}^{N} (y_i \; - \; \hat{y}_i)^2 / N} \tag{1}$$

In both metrics, lower values result in better predictive models. However, the RM SE is more sensitive to high errors. Another possibility to compare regression models is the *Regression Error Characteristic* (REC) curve, which plots the error tolerance (x-axis), given in terms of the absolute deviation, versus the percentage of points predicted SVM present theoretical advantages over NN, such as the absence of local minimum the model optimization phase. In SVM regression, the input $x \in \mathcal{R}^A$ is transformed into a high-dimensional feature space, by using a nonlinear mapping. Then, the SVM finds the best linear separating hyperplane in the feature space:

$$\hat{y} = w_0 + \sum_{i=1}^{m} w_i \phi_i(x) \quad (2)$$

Where $\phi_i(x)$ represents a nonlinear transformation, according to the kernel function $K(x, x') = \sum_{i=1}^{m} \phi_i(x)\phi_i(x')$. To estimate the best SVM, the ϵ-insensitive loss function (Figure 4) is often used[6]. In presenting hyper parameters and less numerical difficulty than other kernels such as polynomials and sigmoid by using the popular Kernel Radial Basis Function

$$K(x, x') = exp\left(-\gamma || x - x' ||^2\right), \gamma > 0 \quad (3)$$

The SVM performance is affected by three parameters *: C– a* trade-off between the model complexity and the amount up to which deviations larger than ϵ are to related ; ϵ – the width of the ϵ-insensitive zone; and γ – the parameter of the kernel. Since the search space for the three parameters is high, the C and ϵ values will be set using the heuristics proposed in : $C = 3$ (for standardized inputs) and $\epsilon = 3\hat{\sigma}\sqrt{\frac{\ln(N)}{N}}$, where and $\hat{\sigma}$ is the standard deviation as predicted by a 3-nearest neighbor algorithm.

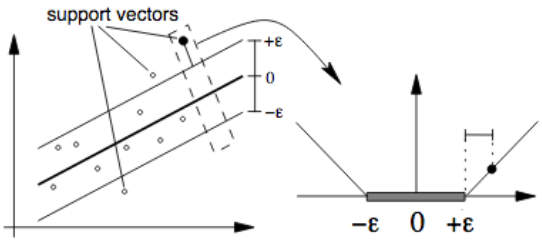
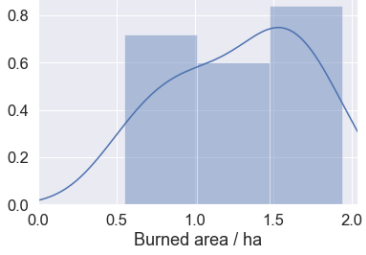


Figure 2. Example of a linear SVM regression and the ϵ-insensitive loss function

3. RESULTS AND DISCUSSION

Predicting fireplace burn of Aek Godang, Northern Sumatera region must assist in directing resources over large areas. A exceptionally interpretable model might provide records on hearth prevention. One may consequently be inclined to seem at multi-linear regression or generalized additive models.

The result of the split the statistics into coaching and trying out sets as shown in Figure 3.



4.

Figure 3. The distribution of the response variable

The response variable burned of Aek Godang, Northern Sumatera area, is extraordinarily skewed towards small fires. It might be beneficial to transform this with e.g. a Log10() scaling. The visual-spatial statistic result of this study proven in Figure four Most fires manifest at central and low X-Y coordinates, with the exception of one very high hearth count grid reference at (8, 6). Comparing complete fires with the total burned region there is some proof that fires at low X are small and numerous, where fires at high X are much less accepted however larger.

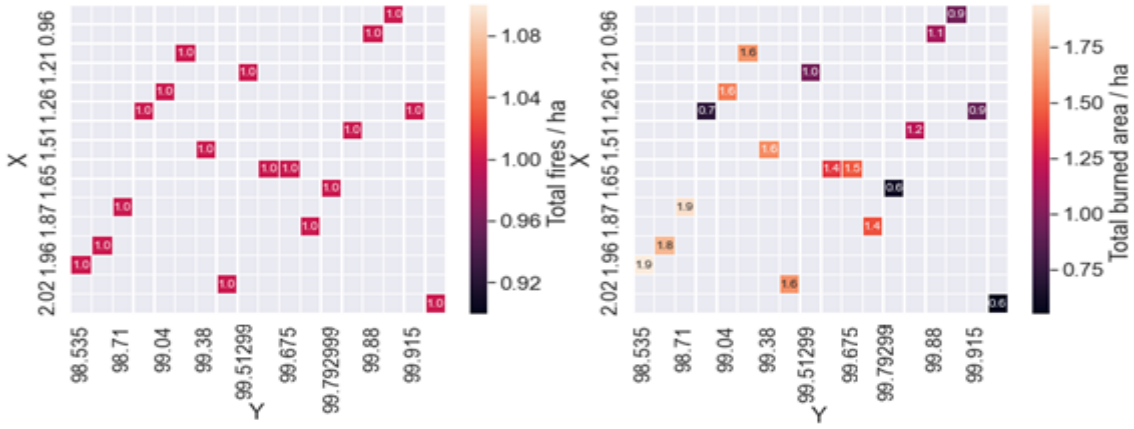


Figure 4. Visualize spatial statistics

The median burned region in Figure 5, reinforces the remaining bullet, that is to say, smaller fires dominate low-X regions the place large fires dominate at high-X regions.

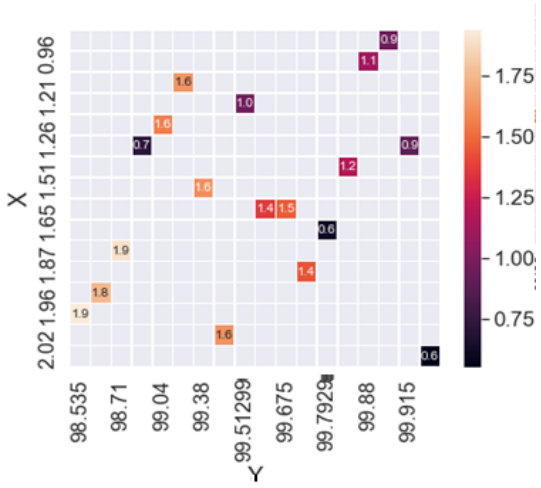


Figure 5. Median Burned

Based on the express the average burned area is biggest in March(Figure 6). However, this may also be the end result of a single or a few fires in view that the width of the distribution is small. The greatest fires have a tendency to occur in the summer time months, Aug via Sep. There is no obvious fashion in location burned on a given day of the week.

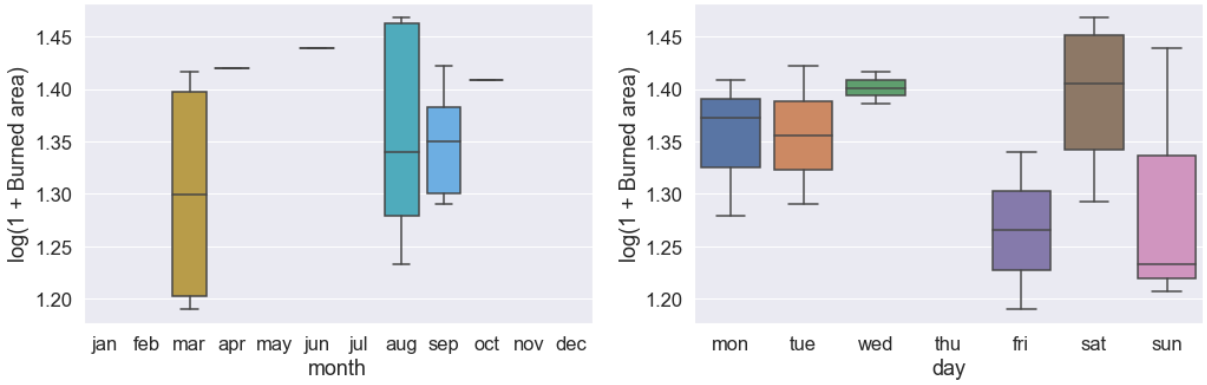


Figure 6. Categorical Variable

In Figure 7, the fire depend as a characteristic of month and day appears like most fires take place in the summer time months of August through September. Most fires manifest on the weekend, possibly pointing to human recreation.

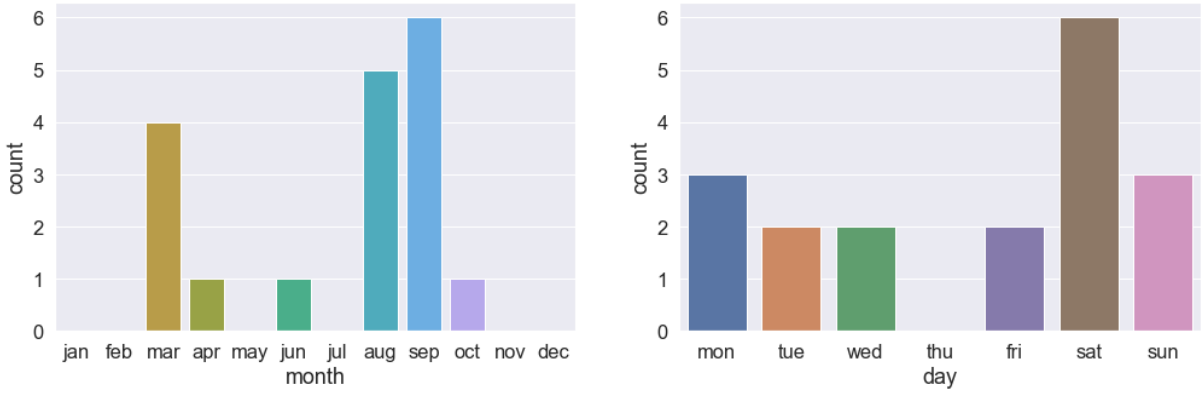


Figure 7. The fire count as a function of month and day look like

The FWI symptoms are all correlated with one another and with temperature. There may also be some (multi) collinearity, which will amplify the variance of a geared up model. It may be beneficial to mix these into a single predictor. We'll stick with the full set of predictors for now.

	X	Y	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
X	1.000000	-0.104349	-0.510968	0.163899	0.173017	-0.445531	-0.379804	0.525578	-0.320112	0.182948	0.152108
Y	-0.104349	1.000000	0.651102	-0.698352	-0.731524	0.603644	0.590436	-0.580939	0.239158	-0.241677	-0.697455
FFMC	-0.510968	0.651102	1.000000	-0.383860	-0.417636	0.966017	0.906448	-0.975714	0.234824	-0.463451	-0.367122
DMC	0.163899	-0.698352	-0.383860	1.000000	0.995787	-0.313409	-0.271251	0.274650	-0.212321	0.241341	0.991549
DC	0.173017	-0.731524	-0.417636	0.995787	1.000000	-0.339549	-0.307984	0.304645	-0.224787	0.251641	0.992462
ISI	-0.445531	0.603644	0.966017	-0.313409	-0.339549	1.000000	0.834487	-0.978134	0.215194	-0.337063	-0.293979
temp	-0.379804	0.590436	0.906448	-0.271251	-0.307984	0.834487	1.000000	-0.862024	0.186511	-0.519856	-0.236786
RH	0.525578	-0.580939	-0.975714	0.274650	0.304645	-0.978134	-0.862024	1.000000	-0.230477	0.425679	0.262224
wind	-0.320112	0.239158	0.234824	-0.212321	-0.224787	0.215194	0.186511	-0.230477	1.000000	-0.195281	-0.215478
rain	0.182948	-0.241677	-0.463451	0.241341	0.251641	-0.337063	-0.519856	0.425679	-0.195281	1.000000	0.216059
area	0.152108	-0.697455	-0.367122	0.991549	0.992462	-0.293979	-0.236786	0.262224	-0.215478	0.216059	1.000000

Figure 8. Correlation matrix

Based on figure 8, the cross-validated imply absolute error from bagging is 0.09. Using default hyperparameters is not the most strong way to examine fashions in this way but we'll assume that the default hyperparameters are set to give practical starting points for most problems. This is very disappointing, the mannequin predicts an nearly regular response. There also appears to be a lower limit on the expected burned area. Does this mirror a lower restriction in the coaching data? It would be prudent to inspect this further.

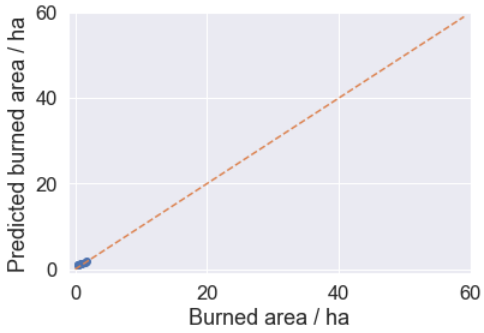


Figure 9. Test predictions against true burned area

The test set deviance increases beyond ~100 iterations, a clear signal that the model is overfitting. It would have been useful to do this checking out on a separate validation dataset as an alternative of the check set. This would have allowed us to go back and address the overfitting. Unfortunately, this is a very difficult dataset to work within that it is small with few if any predictors nicely correlated with the response. We would likely now not get any reward for similarly reducing the measurement of the coaching set.

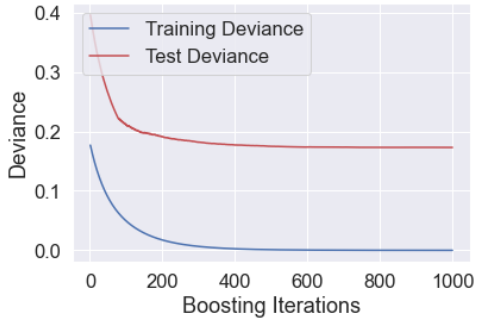


Figure 10. Training and test set deviance

All the fashions give comparable consequences and are tremendously poor, gradient boosting gives the lowest cross-validation error so we will take this forward and attempt to tune the parameters. There is no huge correlation between any one of the predictors and the response. A multilinear regression or generalized additive model is probably no longer going to eke out a signal. Highly nonlinear techniques might be better suited at the rate of interpretation.

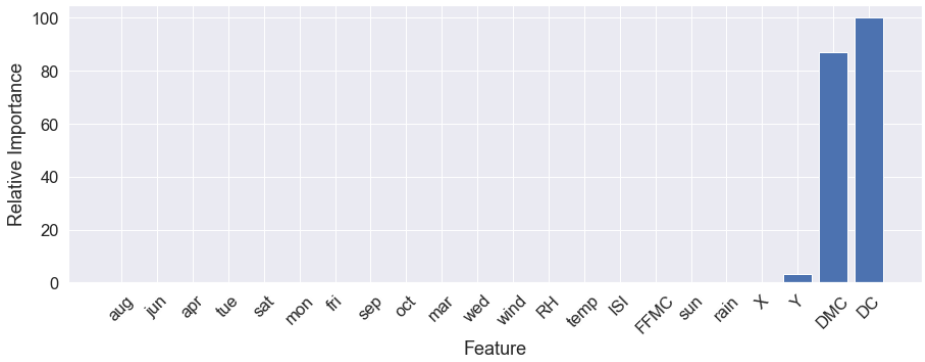


Figure 11. Feature importance

Many of the points have little or no significance in the closing model, they are probable adding noise, indicating that some feature selection might be prudent. The temperature has the easiest significance of all the features, which makes a lot of sense. However, each wind and rain have no importance. One might have expected fires to burn much less vicinity at instances of high precipitation and for high winds to fan the flames.

The wooded area fires dataset used to be presented in Cortez and Morais 2007 [6], the place the authors current a answer to this trouble the use of a trained support vector machine. In assessing the accuracy of their mannequin they produce a REC curve, which plots the error tolerance (x-axis), given in terms of the absolute deviation, versus the percentage of points envisioned in the tolerance (y-axis). The ideal regressor should existing a REC region close to 0.5.

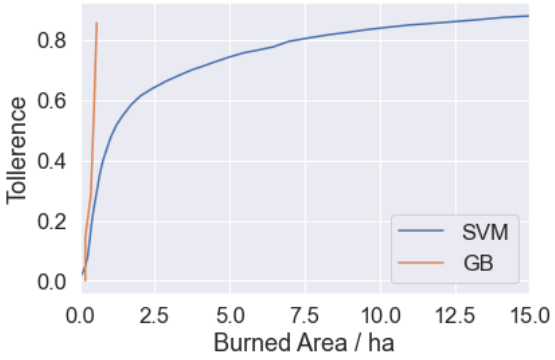


Figure 12. REC Curve

5. CONCLUSION

Based on the result, there is little correlation between the predictor variables and the response variable. It would be prudent to attempt and understand this lack of correlation better. Fire hostilities and prevention exercise may be one motive for the found lack of correlation. The truth that this dataset exists suggests that there is already some effort going into hearth prevention. One ought to imagine a situation in which many fires have the ability to emerge as very massive but are extinguished before they have the risk to do so. If statistics pertaining to furnace prevention is reachable it would likely be an extraordinarily precious addition to this dataset. It was once shown that the gradient boosting model used to be likely overfitting. Controlling the depth of timber and studying charge are two methods which were used to stop overfitting. Scikit-learn gives numerous more, which includes the capacity to enforce a decrease bound on the number of samples in a leaf. This limits the ability of the boosting algorithm to structure leaves that seize single outlying data points, hence decreasing variance and overfitting. As with random forests, introducing randomization into the boosting algorithm can additionally minimize variance. Scikit-learn affords two methods. First by using developing each tree with a random subsample of the education set and 2nd via randomly subsampling the points viewed for each node. In summary, a whole lot greater tuning of the mannequin is possible.

Gradient boosting based totally on the cross-validated mean absolute error from tuned gradient boosting is 0.07, it performs characteristic selection naturally. However, with the use of a validation set, it would have been feasible to use the feature importance plot above to do some guide characteristic selection. In particular, most of the days and months have no relevance to the problem and are probably simply including noise. Unfortunately, the use of a validation set for this motive would always reduce the coaching data, in addition contributing to the situation of making an attempt to eke out a susceptible sign from a small dataset..

REFERENCES

[1] BMKG, “Badan Meteorologi, Klimatologi dan Geofisika,” 2020. .
[2] LAPAN, “Lembaga Penerbangan dan Antariksa Nasional,” 2020. .
[3] D. P. K. H. PKHL, “SiPongi Karhutla Monitoring Sistem,” Jakarta, 2019.
[4] D. Bidang and P. Jauh, *INFORMASI TITIK PANAS (HOTSPOT) KEBAKARAN HUTAN / LAHAN*. 2016.

[5] Nrcan, "Canadian Wildland Fire Information System | Canadian Forest Fire Weather Index (FWI) System," *https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi*, 2020. .

[6] P. Cortez and A. Morais, "A Data Mining Approach to Predict Forest Fires using Meteorological Data," *Proc. 13th Port. Conf. Artif. Intell.*, no. January 2007, pp. 512–523, 2007.