Citation Request:

This dataset is public available for research. The details are described in [Cortez and Morais, 2007].

Please include this citation if you plan to use this database:

P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data.

In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence,

Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December,

Guimaraes, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9.

Available at: http://www.dsi.uminho.pt/~pcortez/fires.pdf

1. Title: Forest Fires

2. Sources

Created by: Paulo Cortez and An�bal Morais (Univ. Minho) @ 2007

3. Past Usage:

P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data.

In Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence,

December, 2007. (http://www.dsi.uminho.pt/~pcortez/fires.pdf)

In the above reference, the output "area" was first transformed with a ln(x+1) function.

Then, several Data Mining methods were applied. After fitting the models, the outputs were

post-processed with the inverse of the ln(x+1) transform. Four different input setups were

used. The experiments were conducted using a 10-fold (cross-validation) x 30 runs. Two

regression metrics were measured: MAD and RMSE. A Gaussian support vector machine (SVM) fed

with only 4 direct weather conditions (temp, RH, wind and rain) obtained the best MAD value:

12.71 +- 0.01 (mean and confidence interval within 95% using a t-student distribution). The

best RMSE was attained by the naive mean predictor. An analysis to the regression error curve

(REC) shows that the SVM model predicts more examples within a lower admitted error. In effect,

the SVM model predicts better small fires, which are the majority.

4. Relevant Information:

This is a very difficult regression task. It can be used to test regression methods. Also,

it could be used to test outlier detection methods, since it is not clear how many outliers

are there. Yet, the number of examples of fires with a large burned area is very small.

5. Number of Instances: 517

6. Number of Attributes: 12 + output attribute

Note: several of the attributes may be correlated, thus it makes sense to apply some sort of

feature selection.

7. Attribute information:

For more information, read [Cortez and Morais, 2007].

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9

2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9

3. month - month of the year: "jan" to "dec"

4. day - day of the week: "mon" to "sun"

5. FFMC - FFMC index from the FWI system: 18.7 to 96.20

6. DMC - DMC index from the FWI system: 1.1 to 291.3

7. DC - DC index from the FWI system: 7.9 to 860.6

8. ISI - ISI index from the FWI system: 0.0 to 56.10

9. temp - temperature in Celsius degrees: 2.2 to 33.30

10. RH - relative humidity in %: 15.0 to 100

11. wind - wind speed in km/h: 0.40 to 9.40

12. rain - outside rain in mm/m2 : 0.0 to 6.4

13. area - the burned area of the forest (in ha): 0.00 to 1090.84

(this output variable is very skewed towards 0.0, thus it may make

sense to model with the logarithm transform).

8. Missing Attribute Values: None