Regression Models
Forest Fire Dataset UCI Machine Learning Repository

Walter R. Vives Castro. *Data Engineering. Universidad Politécnica de Yucatán (*Licenciatura, Km. 4.5. Mérida-Tetiz Tablaje. CP 97357 Ucú,Yucatán, México*)*

Abstract—**This paper will analyze four regression models. Multiple linear regression, multiple linear regression with regularization: lasso, ridge and ElasticNet. The dataset "Forest Fires" is retrieved from UCI Machine Learning Repository. The aim is comparing the regression models and select the best one.**

I. Introduction

Fire is the major stand-renewing disturbance in the circumboreal forest. Weather and climate are the most important factors influencing fire activity and these factors are changing due to human-caused climate change. [1]

"Nearly 85 percent of wildland fires in the United States are caused by humans. Human-caused fires result from campfires left unattended, the burning of debris, equipment use and malfunctions, negligently discarded cigarettes, and intentional acts of arson."

Retrieved from: *2000-2017 data based on Wildland Fire Management Information (WFMI) and* U.S. Forest Service Research Data Archive

Forest fires (also called wildfires), which affect forest preservation, create economical and ecological damage and cause human suffering. Such phenomenon is due to multiple causes (e.g. human negligence and lightnings) and despite an increasing of state expenses to control this disaster, each year millions of forest hectares (ha) are destroyed all around the world. [2] The most common direct human causes of wildfire ignition include arson, discarded cigarettes, power-lines arcs (as detected by arc mapping), and sparks from equipment.

In particular, Portugal is highly affected by forest fires [3]. From 1980 to 2005, over 2.7 million ha of forest area (equivalent to the Albania land area) have been destroyed. The 2003 and 2005 fire seasons were especially dramatic, affecting 4.6% and 3.1% of the territory, with 21 and 18 human deaths.

Therefore, the use of regression models to predict burned areas of forest fires can help to have a better understand of this behavior, either it was caused by human activity or another cause. A key element for a successful firefighting is the fast detection. Thus, having a well-structured machine learning model can help to have a better understanding of the situation and take in consideration the results of the models to avoid forest fires.

II. Data Understanding

Forest Fire Data Set (UCI Machine Learning Repository)

This dataset is public available for research.

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

Data Set retrieved from: Forest_Fire_Data_Set

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.

- Number of Instances: 517
- 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.

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).

The dataset consists of meteorological (and other data) from a northeast region of Portugal. **The aim of this project is to predict the burned area of forest fires.**

The FWI System depends solely on weather readings. Resulting fuel moisture codes and fire behavior indices are based on a single standard fuel type that can be described as a generalized pine forest, most nearly jack pine and lodgepole pine.

The Fire Weather Index System calls for weather observations to be collected from a standard observation site and time. Location standards can be found in the, *Weather Guide for the Canadian Forest Fire Danger Rating System* (Lawson and Armitage, 2008). The system calls for observations to be taken at solar noon, when the sun is at its peak directly overhead.

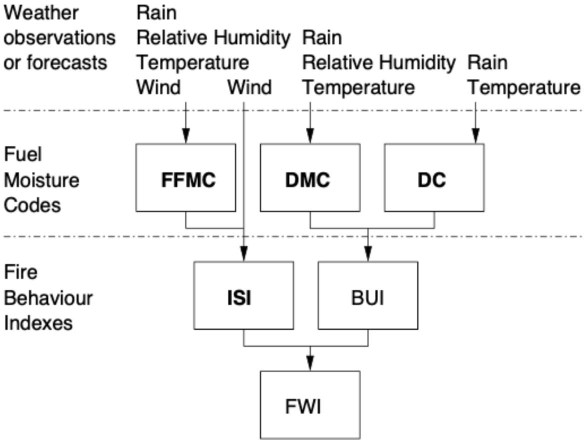


FIGURE 1. The Fire Weather Index structure. From A Data Mining Approach to Predict Forest Fires using Meteorological Data, by P. Cortez, A. Morais.

This study will consider forest fire data from the Montesinho natural park, from the Tra´s-os-Montes northeast region of Portugal (Figure 2). This park contains a high flora and fauna diversity. Inserted within a supra-Mediterranean climate, the average annual temperature is within the range 8 to 12◦C. [2]

III. PREVIOUS WORK

Meteorological data has been incorporated into numerical indices, which are used for prevention (e.g., warning the public of a fire danger) and to support fire man- agement decisions (e.g., level of readiness, prioritizing targets or evaluating guidelines for safe firefighting).

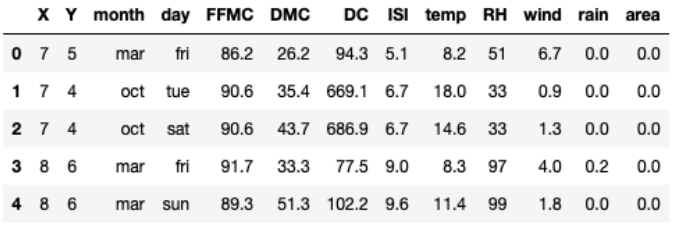
Several DM techniques have been applied to the fire detection domain. For example, Vega-Garcia et al. [4] adopted Neural Networks (NN) to predict human- caused wildfire occurrence. Infrared scanners and NN were combined in to reduce forest fire false alarms with a 90% success. A spatial clustering (FASTCiD) was adopted by Hsu et al. [5] to detect forest fire spots in satellite images.

In 2005 [6], satellite images from North America forest fires were fed into a Support Vector Machine (SVM), which obtained a 75% accuracy at finding smoke at the 1.1-km pixel level. Stojanova et al. [3] have applied Logistic Regression, Random Forest (RF) and Decision Trees (DT) to detect fire occurrence in the Slovenian forests, using both satellite-based and meteorological data. The best model was obtained by a bagging DT, with an overall 80% accuracy.

A novel DM forest fire approach, where the emphasis is the use of real-time and non-costly meteorological data. It was used recent real-world data, collected from the northeast region of Portugal, with the aim of predicting the burned area (or size) of forest fires. Several experiments were carried out by considering five DM techniques (i.e. multiple regression, DT, RF, NN and SVM) and four feature selection setups (i.e. using spatial, temporal, the FWI system and meteorological data).

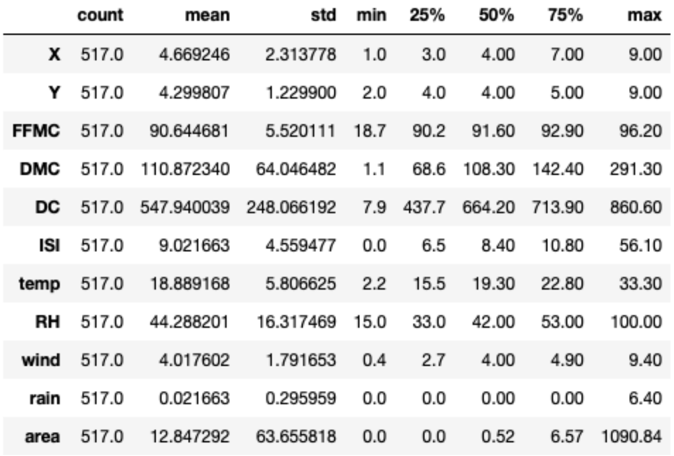
The proposed solution included only four weather variables (i.e. rain, wind, temperature and humidity) in conjunction with a SVM and it is capable of predicting the burned area of small fires, which constitute the majority of the fire occurrences. Such knowledge is particularly useful for fire management decision support (e.g. resource planning). [2]

IV. DATA PREPROCESSING



	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
0	7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0.0	0.0
1	7	4	oct	tue	90.6	35.4	669.1	6.7	18.0	33	0.9	0.0	0.0
2	7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0.0	0.0
3	8	6	mar	fri	91.7	33.3	77.5	9.0	8.3	97	4.0	0.2	0.0
4	8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0.0	0.0

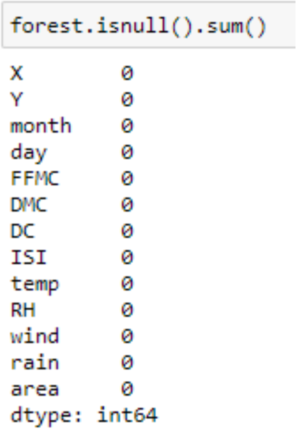
FIGURE 2. Raw Data



	count	mean	std	min	25%	50%	75%	max
X	517.0	4.669246	2.313778	1.0	3.0	4.00	7.00	9.00
Y	517.0	4.299807	1.229900	2.0	4.0	4.00	5.00	9.00
FFMC	517.0	90.644681	5.520111	18.7	90.2	91.60	92.90	96.20
DMC	517.0	110.872340	64.046482	1.1	68.6	108.30	142.40	291.30
DC	517.0	547.940039	248.066192	7.9	437.7	664.20	713.90	860.60
ISI	517.0	9.021663	4.559477	0.0	6.5	8.40	10.80	56.10
temp	517.0	18.889168	5.806625	2.2	15.5	19.30	22.80	33.30
RH	517.0	44.288201	16.317469	15.0	33.0	42.00	53.00	100.00
wind	517.0	4.017602	1.791653	0.4	2.7	4.00	4.90	9.40
rain	517.0	0.021663	0.295959	0.0	0.0	0.00	0.00	6.40
area	517.0	12.847292	63.655818	0.0	0.0	0.52	6.57	1090.84

FIGURE 3. Statistical Descriptors

It can be noticed that there are big differences between the min and max value. Thus, the data must be normalized to reduce that differences between that values.

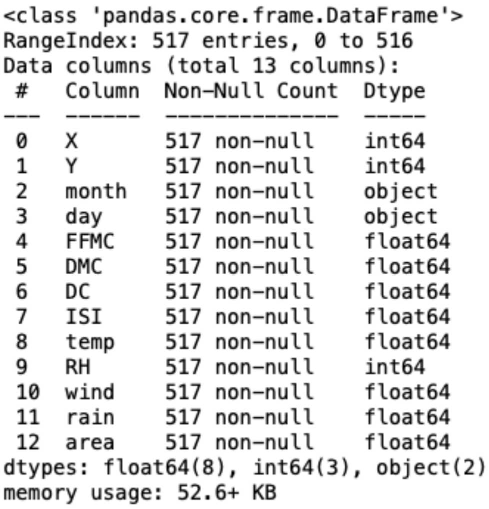


```
forest.isnull().sum()
```

```
X        0
Y        0
month    0
day      0
FFMC     0
DMC      0
DC       0
ISI      0
temp     0
RH       0
wind     0
rain     0
area     0
dtype: int64
```

FIGURE 4. Missing Values

There are not missing values in the dataset.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 517 entries, 0 to 516
Data columns (total 13 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   X       517 non-null    int64
 1   Y       517 non-null    int64
 2   month   517 non-null    object
 3   day     517 non-null    object
 4   FFMC    517 non-null    float64
 5   DMC     517 non-null    float64
 6   DC      517 non-null    float64
 7   ISI     517 non-null    float64
 8   temp    517 non-null    float64
 9   RH      517 non-null    int64
 10  wind    517 non-null    float64
 11  rain    517 non-null    float64
 12  area    517 non-null    float64
dtypes: float64(8), int64(3), object(2)
memory usage: 52.6+ KB
```

FIGURE 5. Data types

Most of our values are int64 and float64, and two object type. Nevertheless, the object types can be used in our model, it will depend if the information in that columns are relevant for our model.

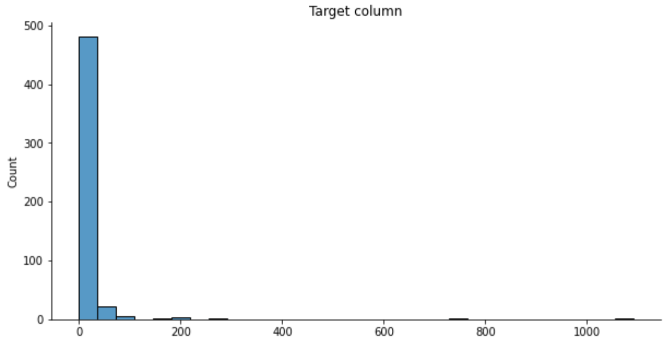


FIGURE 6. Target Feature (area)

It can be noticed that the area feature has a positive skew. The mean and median will be greater than the mode. It can affect our model performance. Therefore, it has to be transformed using a logarithm transform to reduce the positive skew.

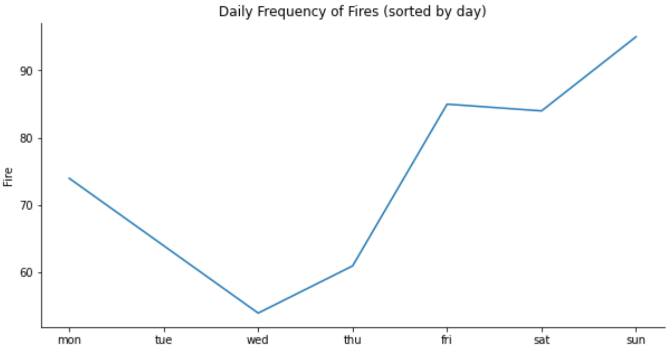


FIGURE 7. Fires by Day

It can be noticed that most of the Fires are in the weekend, that could mean that most of the fires are made by human activity.

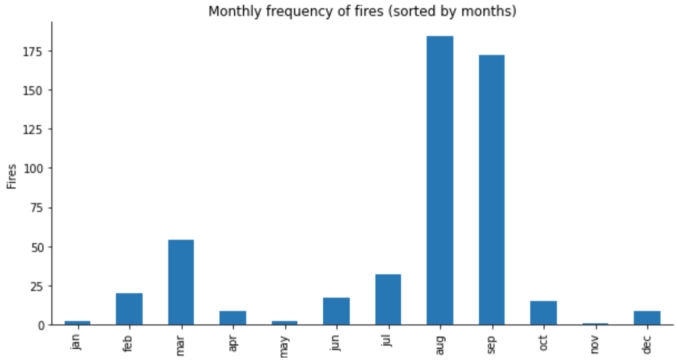


FIGURE 8. Fires by Month

It can be noticed that most of the Fires are in August and September, summer and autumn period. The summer is a hot season, for that reason it can be an important factor that should be periodically monitored.

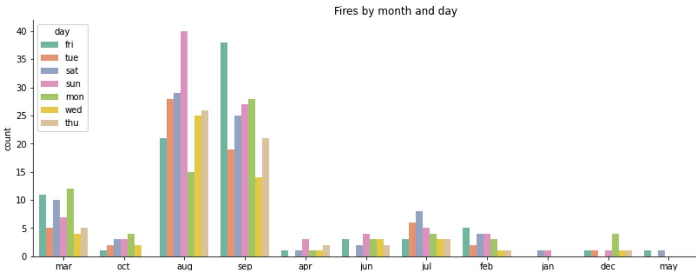
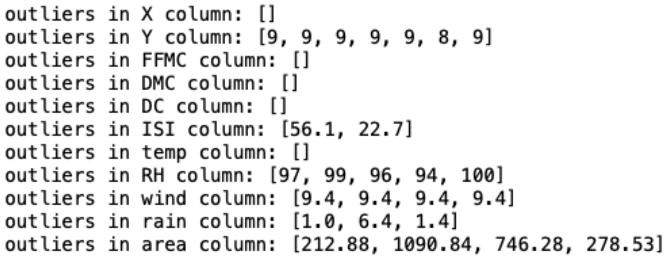


FIGURE 9. Fires by Month and Day

In general, I can be seen that most of the fires are in August and September, in the Fridays and Sundays respectively.



```
outliers in X column: []
outliers in Y column: [9, 9, 9, 9, 9, 8, 9]
outliers in FFMC column: []
outliers in DMC column: []
outliers in DC column: []
outliers in ISI column: [56.1, 22.7]
outliers in temp column: []
outliers in RH column: [97, 99, 96, 94, 100]
outliers in wind column: [9.4, 9.4, 9.4, 9.4]
outliers in rain column: [1.0, 6.4, 1.4]
outliers in area column: [212.88, 1090.84, 746.28, 278.53]
```

FIGURE 10. Outlier Analysis

Given that in the dataset there is not enough records, the outliers will not be deleted. Maybe the outliers of the area column can be removed in order to improve our model.

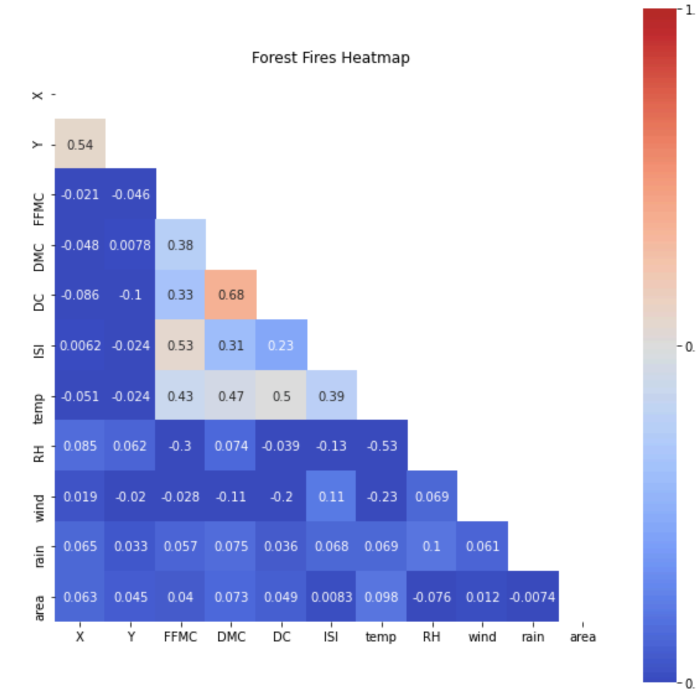


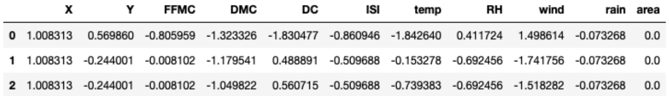
FIGURE 11. Heatmap

It can be notice that the **FFMC** feature has a important correlation with **ISI** given that FFMC is a branch of ISI in the FWI System.

There is a clearly correlation between **DMC** & **DC** features. The reason is that these variables are in the same level in the FWI System. Therefore, this is a warning for us to avoid using both in our model, this can cause Multicollinearity and affect our model.

The features **DMC** & **DC** have correlation with temp feature given that in DMC (Duff Moisture Code) are taking into account three variables: rain, temperature and realtive humidity, and in DC (Drought Code) the variables are rain and temperature.

To determine which features are the best, They have to be compared individually with the target feature and choose the ones that have high correlation. However, the analysis of each independient variable has to be done to avoid multicollinearity.



	X	Y	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
0	1.008313	0.569860	-0.805959	-1.323326	-1.830477	-0.860946	-1.842640	0.411724	1.498614	-0.073268	0.0
1	1.008313	-0.244001	-0.008102	-1.179541	0.488891	-0.509688	-0.153278	-0.692456	-1.741756	-0.073268	0.0
2	1.008313	-0.244001	-0.008102	-1.049822	0.560715	-0.509688	-0.739383	-0.692456	-1.518282	-0.073268	0.0

FIGURE 12. Standardized Features

Standardize features (input variables) by removing the mean and scaling to unit variance.

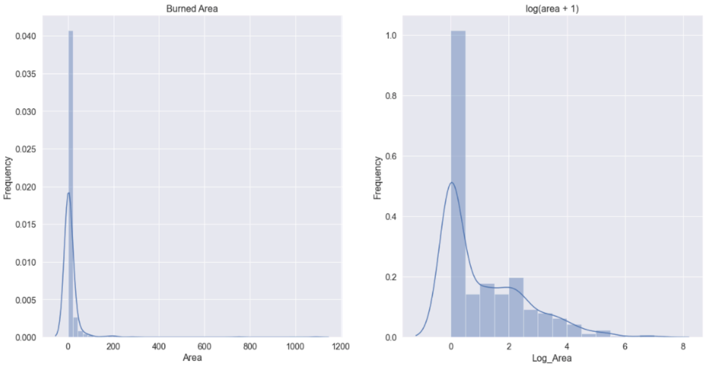
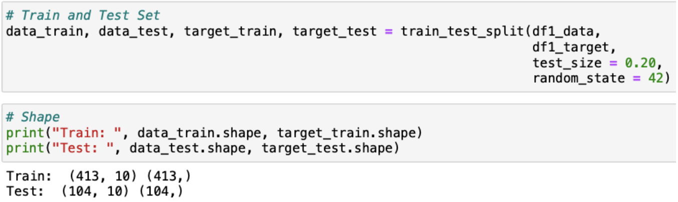


FIGURE 13. Compare the target feature after log transform

The area feature has a highly positive skewed distribution. Therefore, a log transformation had to be done to reduce this problem.

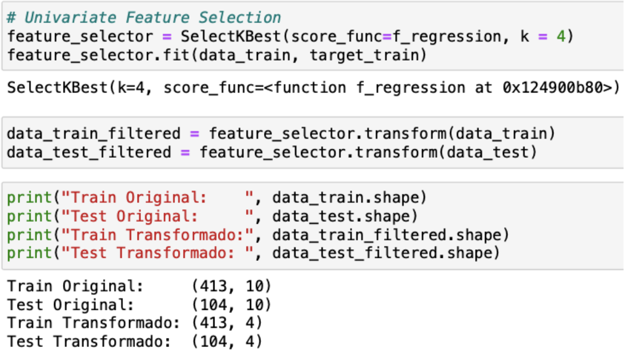


```
# Train and Test Set
data_train, data_test, target_train, target_test = train_test_split(df1_data,
                                                                    df1_target,
                                                                    test_size = 0.20,
                                                                    random_state = 42)
```

```
# Shape
print("Train: ", data_train.shape, target_train.shape)
print("Test: ", data_test.shape, target_test.shape)
```

```
Train:  (413, 10) (413,)
Test:  (104, 10) (104,)
```

FIGURE 14. Train and Test Set



```
# Univariate Feature Selection
feature_selector = SelectKBest(score_func=f_regression, k = 4)
feature_selector.fit(data_train, target_train)
```

```
SelectKBest(k=4, score_func=<function f_regression at 0x124900b80>)
```

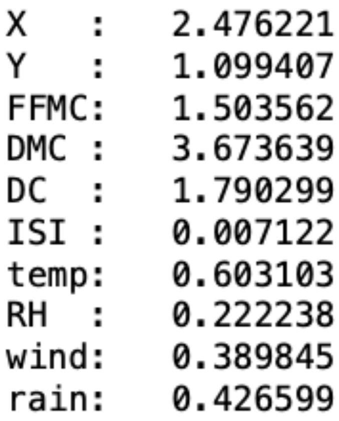
```
data_train_filtered = feature_selector.transform(data_train)
data_test_filtered = feature_selector.transform(data_test)
```

```
print("Train Original:     ", data_train.shape)
print("Test Original:      ", data_test.shape)
print("Train Transformado:", data_train_filtered.shape)
print("Test Transformado: ", data_test_filtered.shape)
```

```
Train Original:     (413, 10)
Test Original:      (104, 10)
Train Transformado: (413, 4)
Test Transformado:  (104, 4)
```

FIGURE 15. Feature Selection

SelectKBest from sklearn library was used using f_regression as our score function. Our K was 4. It was used that number base on the previous work. After that, it will be done an independent vs independent variables and dependent vs independent variables. This will be made in order to avoid multicollinearity and avoid using features that does not apport to our model.



```
X   :  2.476221
Y   :  1.099407
FFMC:  1.503562
DMC :  3.673639
DC  :  1.790299
ISI :  0.007122
temp:  0.603103
RH  :  0.222238
wind:  0.389845
rain:  0.426599
```

FIGURE 16. Score Values

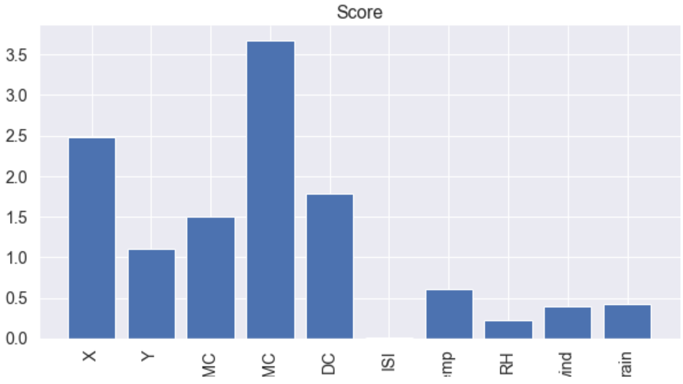


FIGURE 17. Bar plot Score

The scores are made in two steps:

1. The correlation between each regressor and the target.
2. It is converted to an F score then to a p-value

It can be notice that 5 of 10 are less than 1 and 5 of 10 are greater or equal than 1. In our univariate feature selection, using SelectKBest from sklearn, It was stablished k = 4.

Looking the results, It can be noticed that the selected columns are X, FFMC, DMC, DC.

I consider that to improve the performance of the model the geographical vector X (also Y but it was not selected) should not be used in our model, given that the feature do not apport to the model. However, it will be used in order to see its behavior.

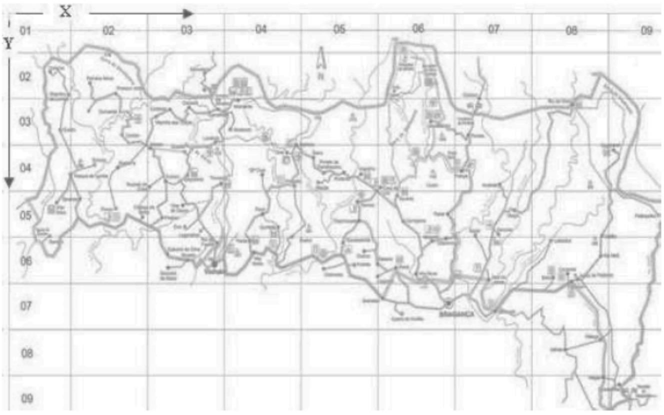


FIGURE 18. The map of the Montesinho natural park. From A Data Mining Approach to Predict Forest Fires using Meteorological Data, by P. Cortez, A. Morais.

Dependent Variable vs Independent Variable Analysis

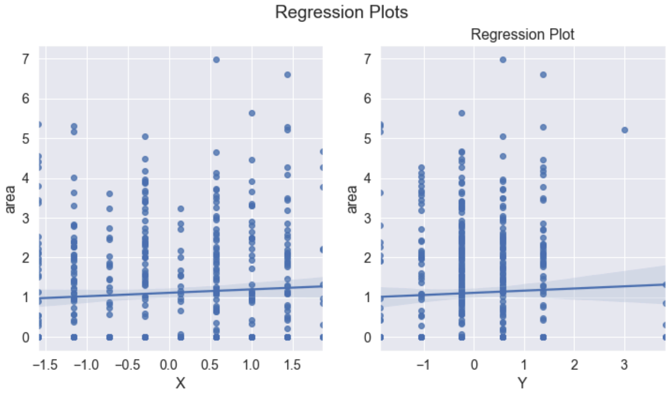


FIGURE 19. Regression Plots (X & Y)

Pearson's correlation coefficient of X & area = 0.061994908322465014

Pearson's correlation coefficient of Y & area = 0.03883821346536808

It can be noticed that the is a weak or non-linear relation between those variables.

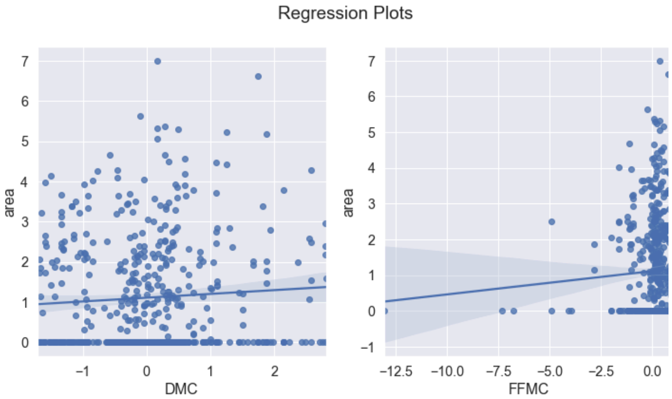


FIGURE 20. Regression Plots (DMC & FFMC)

Pearson's correlation coefficient of DMC & area = 0.06715273981504079

Pearson's correlation coefficient of FFMC & area = 0.046798563676477424

It can be noticed that the is a weak or non-linear relation between those variables.

Even though that **DMC** & **FFMC** had the best scores using f_regression. There is not a clear correlation with our dependent variable area. It can be a problem in our model.

Independent Variable vs Independent Variable Analysis

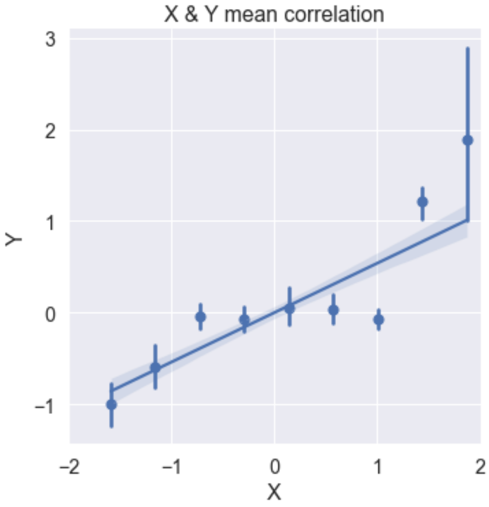


FIGURE 21. X & Y Mean Correlation

It can be notice that there exists a possible correlation given their mean of the features **X & Y**. It means, that both variables should not be taken into account in our model.

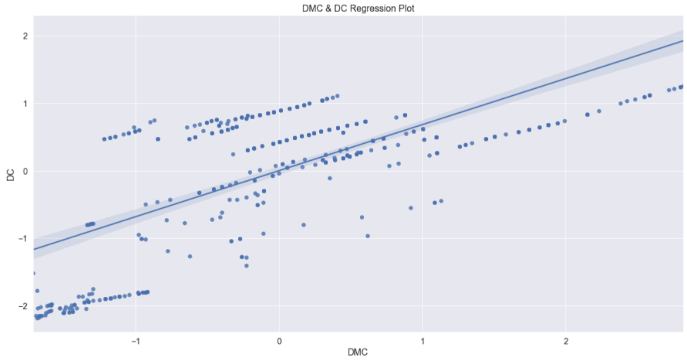


FIGURE 22. DMC & DC Regression Plot

Pearson's correlation coefficient of DMC & DC = 0.6821916119833171

The forest Fire Weather Index (FWI) Canadian system for ratings says that DMC AND DC are in the same classification level (Fuel Moisture Codes). For that reason, there is probability that those features are correlated.

Seeing the result of Pearson's correlation coeficient of those features, it can be concluded that those features that are in the same level (FWI System) are highly correlated.

- If:
  - Negative correlation: -1 <= ρ < 0
  - Neutral correlation: ρ = 0
  - Positive correlation: 0 < ρ <= 1

Our ρ is = 0.682. Therefore, the correlation is strong enough to say that there is a correlation.

For that reason, there must be taken only one feature in our model to avoid **Multicollinearity**. In other words, there could be redundant information. The precision of the estimate coefficients can be reduced given that they are sensitive to small changes.

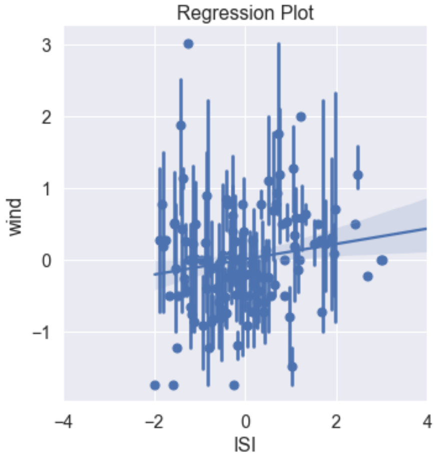


FIGURE 23. Wind-ISI Regression plot

Pearson's correlation coefficient of DMC & DC: 0.10682588792335049

The features wind and ISI have zero/neutral correlation. Even though that in ISI feature the wind is included. Maybe removing the outliers can improve that correlation.

Final Features Selected

The features selected were['X', 'FFMC', 'DMC', 'DC']. However, in order to improve the model, the DC will not be take into account in our model given that has high correlation with DMCfeature and it can cause multicollinearity.

Therefore, the final features selected are ['X', 'FFMC', 'DMC'].

Our final data train and data train are:

- Data train: (413, 3)
- Data test: (104, 3)

V. DATA MODELLING

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

FIGURE 24. R Square Formula

R square: The ratio of the regression error against the total error tells you how much of the total error remains in your regression model. Subtracting that ratio from 1.0 gives how much error you removed using the regression analysis.

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} | x_i - x |$$

FIGURE 25. Mean Absolute Error (MAE)

The **MAE** measures the average magnitude of the errors in a set of forecasts, without considering their direction.

The **MAE** is a linear **score** which means that all the individual differences are weighted equally in the average.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \hat{x}_i\right)^2}{N}}$$

RMSD = root-mean-square deviation
i = variable i
N = number of non-missing data points
x_i = actual observations time series
$\hat{x}_i$ = estimated time series

FIGURE 26. Root-Mean Squared Error/Deviation

1. **Multiple Linear Regression:**

Intercept:
w0 = 1.10

Coefficient:
w1 = 0.11
w2 = 0.04
w3 = 0.11

Evaluation:

MSE_sklearn = 2.238
MAE_sklearn = 1.211
RMSE_sklearn = 1.496
R2_sklearn = -0.018
RMSE on 5-fold Cross Validation: 0.00

With the statistical method K-fold was obtained RME = 0.00. This means, that with the resampling, with k = 5 was obtained a better performance in our model.

Our R square was **negative**, that means that our model is worst than using just the mean. In this case is better use the mean value.

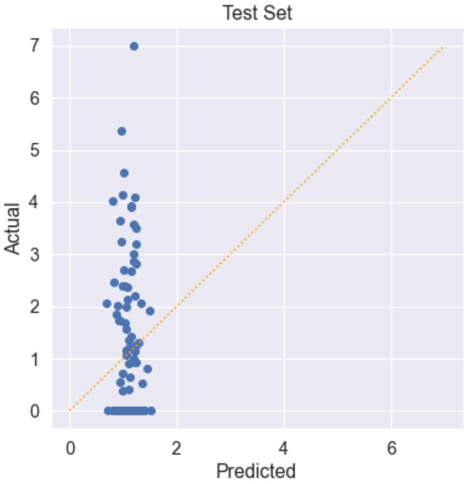


FIGURE 26. Prediction vs Actual Value

It can be clearly noticed that the prediction of our model was not good enough.

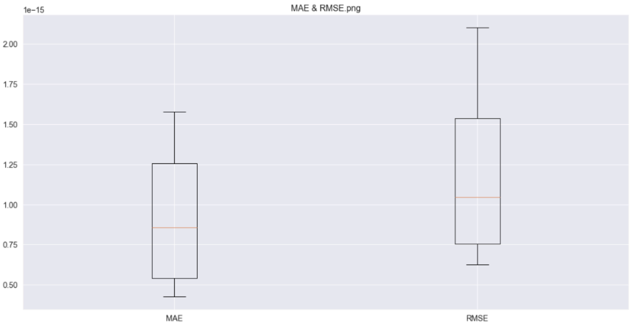


FIGURE 27. MAE & RMSE Boxplot

RMSE gives a relatively high weight to large errors. Mean Absolute Error (MAE): MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

As It can be seen, the RMSE is higher than MAE, because RMSE is relatively high weigh to large errors, given that the error are squared before they are averaged. So, RMSE can be useful when large errors are undesirable.

2. **Multiple linear regression with Lasso Regularization:**

   Evaluation:

   MSE_sklearn = 2.199
   MAE_sklearn = 1.203
   RMSE_sklearn = 1.483
   R2_sklearn = -0.001

Our R square was **negative**, that means that our model is worst than using just the mean. In this case is better use the mean value.

3. **Multiple linear regression with Ridge Regularization:**

   Evaluation:

   MSE_sklearn = 2.238
   MAE_sklearn = 1.211
   RMSE_sklearn = 1.496
   R2_sklearn = -0.018

4. **Multiple linear regression with ElasticNet Regularization:**

   Evaluation:

   MSE_sklearn = 2.199
   MAE_sklearn = 1.203
   RMSE_sklearn = 1.483
   R2_sklearn = -0.001

VI. CONCLUSION AND FUTURE WORK

The models that had the best results were Multiple Linear Regression: Lasso and ElasticNet with an R2 of -0.001.

Although they had the best results compared to the other models, the results of their R2 were negative, that means that using the mean value we would obtain better values. It can be concluded that these linear regression models are not an efficient way to model the behavior of this data set.

It could use object variables like the month to see if the performance of the model improves. Care must also be taken when making the selection of features because the features have a high correlation, and this can lead to multicollinearity which is redundant information for the model. Also, avoid using independent variables that do not share a correlation with the dependent variable since this would not contribute anything to the performance of the model.

For future work, the ideal would be to test the months column and discard the X and Y columns. Also, some classification technique could be adapted to obtain better performance in the model.

REFERENCES

[1] 1. M.D. Flanigan, B.D Amiro, K.A Logan, B.J Stocks & B.M Wotton. FOREST FIRES AND CLIMATE CHANGE IN THE 21ST CENTURY. Springer 2005.

[2] P. Cortez & A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. Department of Information Systems/R&D Algoritmi Centre, University of Minho, 4800-058 Guimara˜es, Portugal.

[3] European-Commission. Forest Fires in Europe. Technical report, Report N-4/6, 2003/2005.

[4] C. Vega-Garcia, B. Lee, P. Woodard, and S. Titus. Applying neural network technology to human-caused wildfire occurence prediction. *AI Applications*, 10(3):9–18, 1996.

[5] W. Hsu, M. Lee, and J. Zhang. Image Mining: Trends and Developments. *Journal of Intelligent Information Systems*, 19(1):7–23, 2002.

[6] D.Mazzoni, L.Tong,D.Diner,Q.Li,andJ.Logan.UsingMISRandMODISDataForDetection and Analysis of Smoke Plume Injection Heights Over North America During Summer 2004. *AGU Fall Meeting Abstracts*, pages B853+, December 2005.

[7] W. Hsu, M. Lee, and J. Zhang. Image Mining: Trends and Developments. Journal of Intelligent Information Systems, 19(1):7–23, 2002.