A Data Science Approach to Predict Forest Fires

Lina Fahed

1. Problem description and objectives

Forest fires are a major environmental issue, creating economical and ecological damage while endangering human lives. Fast detection is a key element for controlling such phenomenon. To achieve this, one alternative is to use automatic tools based on local sensors, such as provided by meteorological stations. In effect, meteorological conditions (e.g. temperature, wind) are known to influence forest fires and several fire indexes, such as the forest Fire Weather Index (FWI), use such data. A detailed description of the forest fires and the data science problem have been published in this paper [Cortez and Morais, 2007][1]: http://www3.dsi.uminho.pt/pcortez/fires.pdf

The aim of this data science project is to understand the forest fires and to predict the burned area of forests, in the northeast region of **Portugal**, by using meteorological and other data. The objective is to determine what factors impact the most forest fires: when (temporal dimension), where (spatial dimension), what meteorological factors, how “big”, “intense”, “speed” will be the forest fire, etc. Be creative and surprise us with new interesting questions.

2. Data description

Download data, find more about problem and data description here: http://archive.ics.uci.edu/ml/datasets/Forest+Fires

- Number of Instances: 517
- Number of Attributes: 13 attributes

The forest Fire Weather Index (FWI) is the Canadian system for rating fire danger and it includes six components: Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI) and FWI. The first three are related to fuel codes: the FFMC denotes the moisture content surface litter and influences ignition and fire spread, while the DMC and DC represent the moisture content of shallow and deep organic layers, which affect fire intensity. The ISI is a score that correlates with fire velocity spread, while BUI represents the amount of available fuel. The FWI index is an indicator of fire intensity and it combines the two previous components. Although different scales are used for each of the FWI elements, high values suggest more severe burning conditions.

1 [Cortez and Morais, 2007] 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, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. Available at:http://www3.dsi.uminho.pt/pcortez/fires.pdf

Attributes Description:

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

- X : x-axis spatial coordinate within the Montesinho park map: 1 to 9
- Y : y-axis spatial coordinate within the Montesinho park map: 2 to 9
- month : month of the year: 'jan' to 'dec'
- day : day of the week: 'mon' to 'sun'
- FFMC : Fine Fuel Moisture Code index from the FWI system: 18.7 to 96.20
- DMC : Duff Moisture Code index from the FWI system: 1.1 to 291.3
- DC : Drought Code index from the FWI system: 7.9 to 860.6
- ISI : Initial Spread Index index from the FWI system: 0.0 to 56.10
- temp : temperature in Celsius degrees: 2.2 to 33.30
- RH : relative humidity in %: 15.0 to 100
- wind : wind speed in km/h: 0.40 to 9.40
- rain : outside rain in mm/m2: 0.0 to 6.4
- area : the burned area of the forest (in ha): 0.00 to 1090.84 (this output variable is very skewed towards 0.0).

See this link for spatial coordinates: The map of Montesinho natural park

3. Work to be done

Your mission consists in setting up a complete methodology of data science by applying the techniques seen in the course. We propose to follow the following steps (points for each step are indicated, total points / 20)

A. Statistical analysis and feature engineering (3 points)

- Uni- and multi-dimensional statistics: evaluation of data quality (empty, incorrect values, ...), understanding of structure, relationships between variables using regression and other statistical techniques
- Possible re-coding of variables, transformation, creation of new variables, …

B. Un-supervised Learning: Clustering (4 points + Bonus 1 point)

- Use clustering techniques in order to understand more about similar forest fires by analyzing the distribution of forest fires according to different factors: spatial (where), temporal (when), intensity, velocity, etc.….
- **Bonus**: Try to exploit several factors at the same. What new information can you get about the forest fires?

C. **Supervised Learning: Regression problem formulation (4 points + Bonus 1 point)**

- What is your target variable among current variables in the dataset, i.e. the variable to predict? Is it the *burned area*? Discuss this. Notice that, at this stage, we consider the problem as a regression problem, i.e. the target variable is numerical (same as all variables in the dataset are numerical).
- Construct a model in order to predict the burned area of the forest. Try several models, and choose the best one. You can try for example Linear regression, Regression decision trees, ….
- **Bonus**: Can you predict other variables like spatial, and temporal coordinates of the forest fire? If yes, construct a model to do that.

D. **Supervised Learning: Classification problem formulation (5 points)**

This is the creative part of the project.

- We ask you to create a new non-numerical variable (not a ration scale), calculated from the starting already existing variables, that you think will be more representative of the forest fires. You can ask yourself the question: what is the most interesting thing that I need to predict about forest fires (other than the *area in ha*) that allows to understand instantaneously the predicted situation? It can be a new binary variable (e.g., fires (1) or no-fires (0)) or a categorical variable (e.g., small/medium/big fire), etc… You are free to propose a new variable, however you have to discuss your proposition……. One condition to respect: the new variable should not be a ration scale, and should represent data labels.
- This new variable will be the target variable. Thus, the regression problem (seen in the previous paragraph) will be transformed into a classification problem. You will now be able to apply different methods seen in previous courses. Thus, you can propose a model able to predict the forest fires (defined according to your new created target variable). You are asked also to evaluate your model, on a test set using several evaluation measures like accuracy, precision, ……..
- **Bonus**: Try new advanced models

E. **<u>Quality of the code, analysis and understanding</u> (2 points)**

Particular attention will be paid to the quality of your analysis, your ideas on the proposed problem. A very well commented notebook will be the minimum rendering.

Practical information: submission of the project

Teams composition for the project:

- Ishita, Pradyumn, Prashant

Submission Deadline: The project is due by **26 May, at 11:59 pm (23h59) (no deadline extension will be possible)**

How to submit the project?

- You have to submit a compressed file strictly named as: **name1_name2_name3.zip** (where *name1, name2, name3*: are last names of students). Only a **.zip** file is accepted.
- This compressed file **.zip** contains your Python Jupiter Notebook file, dataset, and other files or figures if necessary, ….
- Send the .zip file by e-mail

Project Defense/presentation : May 30

The presentation of your project represents 20% of the total points of the evaluation of this course. During the presentation, your understanding of your Python code and of provided solution and results will be particularly evaluated.

- Your presentation duration: 30 minutes
- Questions/answers: 10 minutes

Some resources

Here are some solutions that you can use and can inspire you. However, if you use them without **understanding** them, it will not be in your interest and will work against you, as we can easily detect it during your presentation and questions. So be aware of the incorrect use of these resources.

- https://www.mdpi.com/2076-3263/10/2/53/htm
- https://www.kaggle.com/code/johndoea/kmeans-forest-fire-clustering/notebook
- https://www.kaggle.com/code/e96031413/forest-fire-area-classification/notebook
- https://www.kaggle.com/code/sazack/forest-fire-burned-area-prediction/notebook
- https://github.com/muralikgs11/forest-fires-svm/blob/master/final_svm.py
- https://www.kaggle.com/elikplim/forest-fires-data-set/code

You can find next, the details about how your project and presentation will be evaluated.