Algeria Forest Fires Dataset Analysis, Prediction and Clustering

Ayman Makhoukhi

Eötvös loránd university, Budapest Pázmány Péter stny. 1/C, 1117, Hungary
ayman.makhoukhi@gmail.com

Abstract. Back in 2012, an estimated 20,000 hectares of land have been ravaged by forest fires in the north of Algeria between June and September. Forest fire is a disaster that leads to serious issues to the affected nation which causes increasing carbon, climate change and threatening the humanity as a whole. So predicting such essential environmental issue is essential to mitigate this threat. during this paper the main purpose is to use Data Science (DS) and Machine Learning models (ML) to predict forest fire outbreak which are presented on this paper and they depend on weather elements and Fire Weather Index (FWI) components.

Keywords: Explatory Data Analysis · Data Mining · Fire Prediction · Classification · Clustering

1 Introduction

According to the National Oceanic and Atmospheric Administration, Warmer temperatures and lower relative humidity make the fuels more receptive to ignition. Stronger winds supply oxygen to fire, and When hot, dry, and windy conditions occur simultaneously, wildfires can spread quickly. The work on this paper aims to create an exploratory data analysis report using Algerian forest fires dataset which help on verifying the missing data and other errors. Get maximum insight into the dataset and its underlying structure etc. In addition to this, creating some visuals that are generally designed to help decision-makers, to set goals and to understand what happened and why something happened and with the same informations appropriate changes could be implemented. Thereby creating predictive machine learning models that classify whether the region is

on or not on fire based on certain attributes. The main idea is proposing a fire prediction based on integrating some data mining techniques for that and since the classification is two classes (Fire & Not Fire), the binary classification is considered. As a first step, performing some analysis on the data by applying some EDA along with some data preprocessing to make the data ready to be fitted into a Machine Learning model as well as creating some data visuals for a better understanding of the dataset, for the second step creating some ML predictive models for the binary classification and compare between their different performance and choosing the best one in order to create a classification report for the chosen model. the last section will be devoted to the clustering of the dataset and interpreting the clusters into low-moderate-high and very high danger fire using the proposed graph of Forest Fire Danger Class Criteria from the National Rural Fire Authority.

2 Explatory Data Analysis

2.1 Dataset

By following the important practices in data preparation including checking the dataset columns, data formats, verifying and fixing data types, fixing the uniqueness of certain attributes by removing unnecessary white-spaces from the headers and its observations values... Ending up with a dataset composed of 244 instances that regroup a data of two regions of Algeria (122 instances for each region), namely the Bejaia region located in the northeast of Algeria and the Sidi Bel-abbes region located in the northwest of Algeria. The period from June 2012 to September 2012. The dataset includes 14 attribues and 1 output attribue (class) that represents if the region is on fire or not on fire.

	day	month	year	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	Classes	Region
121	30	09	2012	25	78	14	1.4	45	1.9	7.5	0.2	2.4	0.1	not fire	1
122	01	06	2012	32	71	12	0.7	57.1	2.5	8.2	0.6	2.8	0.2	not fire	2
123	02	06	2012	30	73	13	4	55.7	2.7	7.8	0.6	2.9	0.2	not fire	2
124	03	06	2012	29	80	14	2	48.7	2.2	7.6	0.3	2.6	0.1	not fire	2
125	04	06	2012	30	64	14	0	79.4	5.2	15.4	2.2	5.6	1	not fire	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
239	26	09	2012	30	65	14	0	85.4	16	44.5	4.5	16.9	6.5	fire	2
240	27	09	2012	28	87	15	4.4	41.1	6.5	8	0.1	6.2	0	not fire	2
241	28	09	2012	27	87	29	0.5	45.9	3.5	7.9	0.4	3.4	0.2	not fire	2
242	29	09	2012	24	54	18	0.1	79.7	4.3	15.2	1.7	5.1	0.7	not fire	2
243	30	09	2012	24	64	15	0.2	67.3	3.8	16.5	1.2	4.8	0.5	not fire	2

Fig. 1. Algeria Forest Fires Dataset.

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), Build up 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.

2.2 Visuals Interpretation

From the below bar plot we can clearly observe that there's no such significant imbalance in the dataset or any extreme disproportion among the number of examples of each class of the problem.the distribution is of 57% 'fire' and 43% 'not fire', such a small imbalance will not create big issues while building and evaluation the Machine Learning models later on in the report.

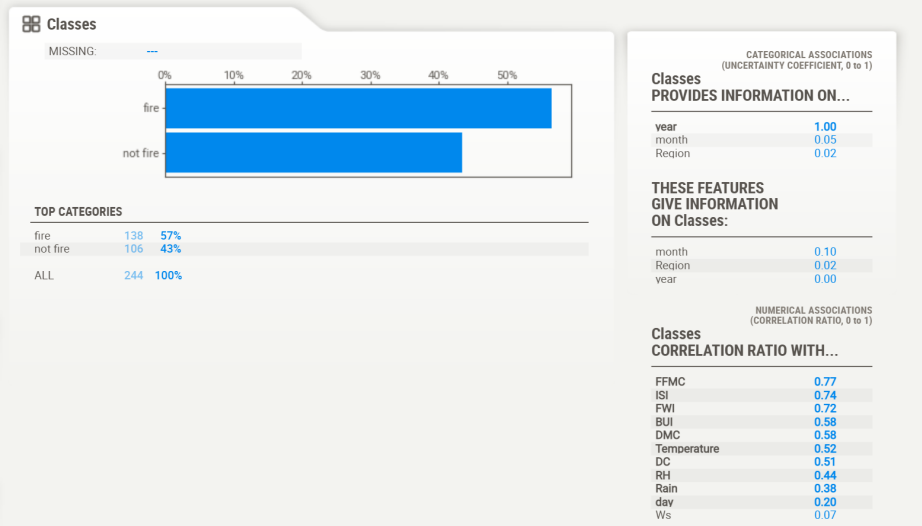

Fig. 2. Dataset Imbalance.

In the right bottom corner of the above fig, we can clearly see that the correlation ratio between the different attributes and the fire outbreak which confirms the indexing of the Canadian system that FWI components has a huge effect on this threat. For example the correlation ratio of FFMC is 77% and ISI is 74% with the fire outbreak.

The following figure represents how some weather metrics lead to the fire outbreak such that climate scientists define summer as June, July and August where we can observe the temperature is highly increasing in that order and also how the wind speed causes this threat as well...

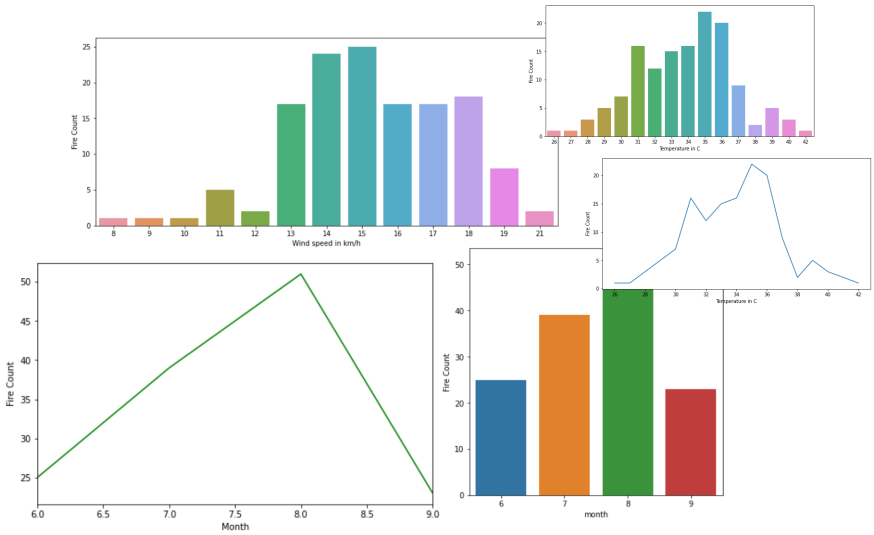

Fig. 3. Dataset Imbalance.

3 Building the Classifier

3.1 Different Classifiers Comparison

Thanks to the sickit learn python library, we will now build Machine Learning models with Some basic hyper-parameters tuning and evaluate their predictions. Those are the models used for this report:
Naive Bayes - Decision Tree - Random Forest - K Nearest Neighbors - Neural Network- Support Vector Classifier.

For a fair comparison we will use Cross-validation since it is usually the preferred method because it gives our models the opportunity to train on multiple train-test splits. This gives us a better indication of how well our model will perform on unseen data. ... That makes the hold-out method score dependent on how the data is split into train and test sets.

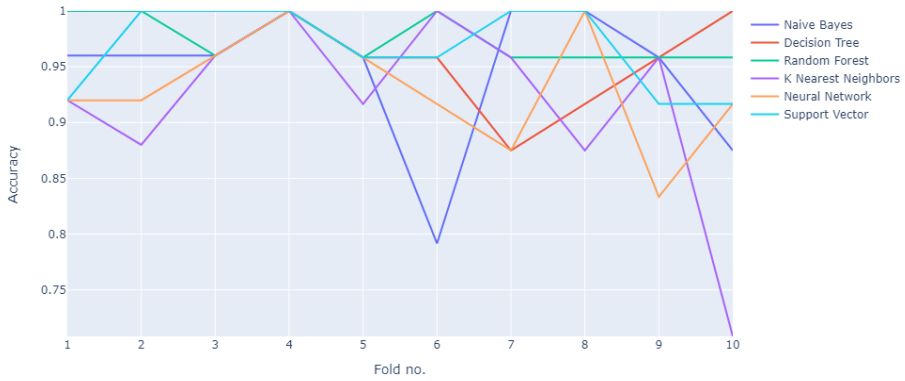

Fig. 4. ML Models Performance.

There are plenty of ways to avoid overfitting while training Machine learning models such as :

- Cross-validation
- Train with more data
- removing irrelevant input features.
- Early stopping that could be implemented with decision tree model by **Pruning** which is a data compression technique in machine learning that reduces the size of decision trees by removing sections of the tree Pruning of decision trees to avoid overfitting that uses cost-complexity pruning by increasing "ccp-alpha" parameter As it increases, more of the tree is pruned.
- Ensemble methods etc.

Since this dataset is pretty small, it is very likely that all the possible features being encountered in the training process so it is quite hard to avoid overfitting for smaller datasets.

3.2 Classification Report

Tree-based classifiers such as DecisionTree and RandomForest as well as the Support Vector Classifier with a linear kernel outperforms the others models.

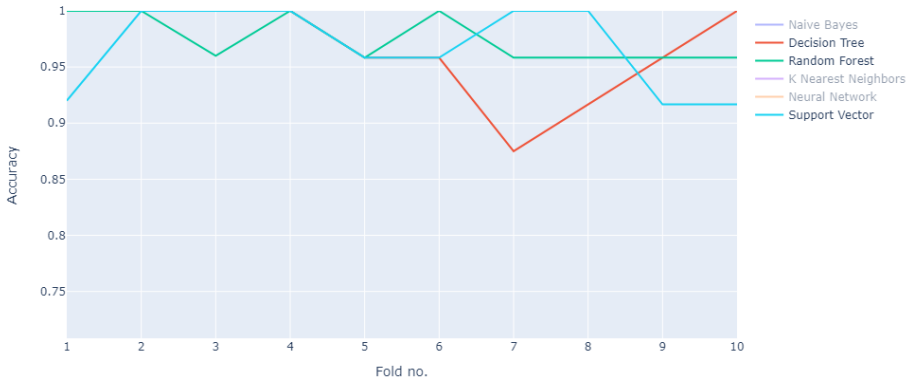

Fig. 5. The models with the highest score

These are actually accurate results and they were kind of predictable, following this handy cheat sheet about details when certain algorithms could be used for different types of machine learning problems.

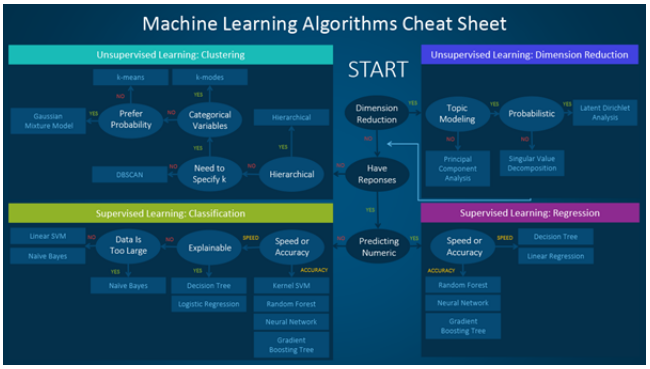

Fig. 6. Machine Learning Algorithms Cheat Sheet.

Since our dataset is small, SVC comes in hands because it is a better suite in case of data with large number of features and lesser observations, and Random forest is used for speed and better accuracy where Decision Tree and Linear SVM are great for speed as well.

We chose the Random Forest Model as our classifier, after fitting the model we can print a report about the performance.

	precision	recall	f1-score	support
0	0.96	1.00	0.98	27
1	1.00	0.98	0.99	47
accuracy			0.99	74
macro avg	0.98	0.99	0.99	74
weighted avg	0.99	0.99	0.99	74

Fig. 7. Random Forest Classification Report

3.3 Drawing The confusion Matrix

Using The confusion matrix as our performance measure for our machine learning classification problem, Random forest has ONLY one Type 2 Error (False Negative) where the model predicted that there is no fire but there actually is.

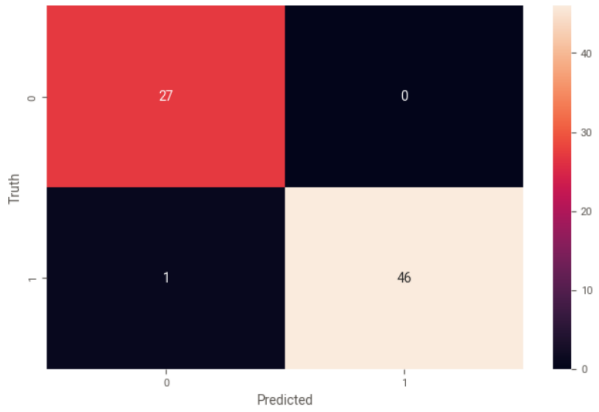

Fig. 8. Random Forest Confusion Matrix

4 Clustering

We will be using **Within-Cluster Sum of Square** that holds the abreviation WCSS which is the sum of squared distance between each point and the centroid in a cluster, By analyzing the graph below we can see that the graph will rapidly change at a point and thus creating an elbow shape. That point corresponds to the optimal K value which is the optimal number of clusters. Our K=3

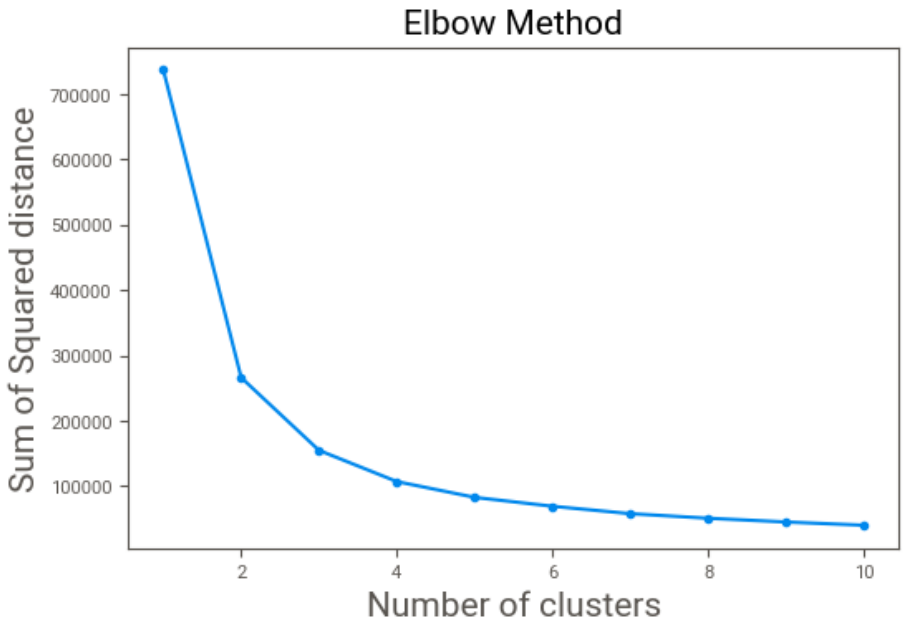

Fig. 9. Elbow Method for the optimal K

4.1 KMeans on the Whole Dataset

In order to visualize our clusters results using a 2D Scatter plot, we will be using a famous unsupervised dimensionality reduction technique namely **PCA** which stands for **Principal Component Analysis** to reduce the number of dimensions .

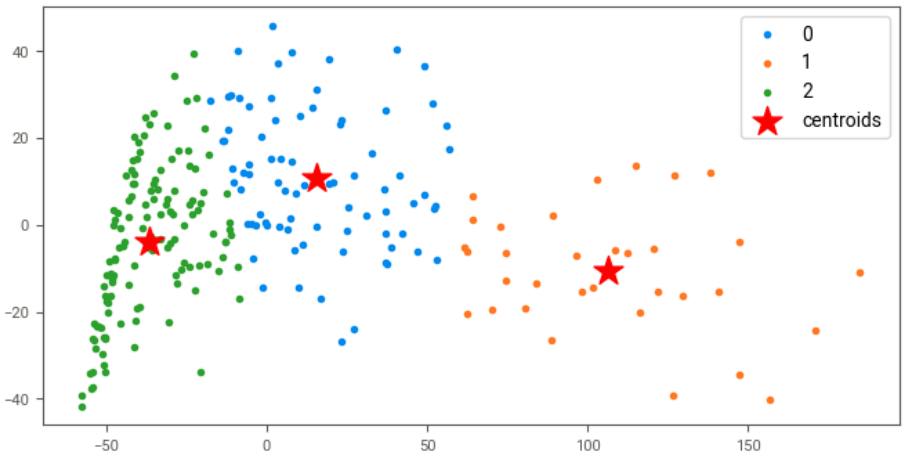

Fig. 10. KMeans on the whole dataset

4.2 KMeans on 2 Features

In order to chose the right features, we need to break down The FWI System: Considering Canadian Forest FWI System we knows that the weather observation such as temperature, relative humidity, wind speed are the main metrics for calculating the fuel moisture codes (FFMC,DMC,DC), these ones are reponsible for changing the numeric rating of the fire behaviors indices (ISI,BUI), these last two indices shapes The FWI System.

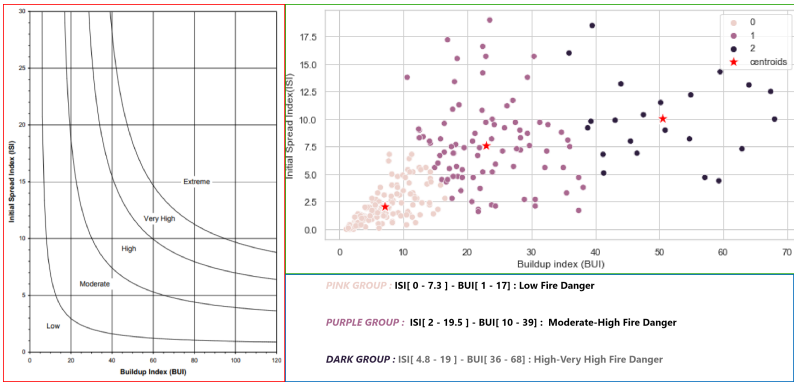

Fig. 11. KMeans on ISI and BUI

5 Frequent Pattern Mining

Some researches show that one limitation of classical association rule mining algorithms that has been addressed is the fact that they require categorical data, they cannot directly deal with numeric attributes. ... It can be used as a pre-processor to "standard" association rule mining algorithms like Apriori. In such cases like these discretization algorithm come in hand.

6 Conclusion

After trying multiple machine learning models to classify whether the region is on fire or not fire in Algeria, we come up with promising results are obtained with the Random Forest Model in terms of performance with about 96% and 98% for the f1-score and the accuracy respectively. But still the accuracy metric does not holds good for imbalanced data. In business scenarios, most data won't be balanced and so accuracy becomes poor measure of evaluation for our classification model, and with a huge and imbalance dataset the over-fitting could affect negatively the model performance. In this case some other machine learning techniques should be implemented to avoid the over-fitting.

References

1. Fire Weather Index System : https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system
2. Algeria Maps of summer 2012 wildfires: https://www.un-spider.org/news-and-events/news/algeria-maps-summer-2012-wildfires-available
3. Canadian FWI System: https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi
4. Europe's eyes on Earth-Fire Danger Forecast: https://effis.jrc.ec.europa.eu/about-effis/technical-background/fire-danger-forecast
5. Martin E. Alexander - Proposed Revision of Fire Danger Class Criteria for Forest and Rural Areas in New Zealand: https://gfmc.online/wp-content/uploads/Proposed-Revision-Fire-Danger-Classes-New-Zealand.pdf
6. https://www.kdnuggets.com/2020/05/guide-choose-right-machine-learning-algorithm.html
7. https://elitedatascience.com/overfitting-in-machine-learning
8. https://elitedatascience.com/dimensionality-reduction-algorithms
9. Relative Unsupervised Discretization for Association Rule Mining : https://link.springer.com/content/pdf/10.1007/3-540-45372-5_15.pdf