A picture containing font, white, symbol, graphics

Description automatically generated

«Algerian Forest Fires Dataset Data Set».

 Alice Anne-Catherine J Destrait (230AIB016)

Professor: Alla Anohina-Naumeca

Report

Practical Assignment 2

<https://github.com/alicedes/machinelearningwithOrange/tree/main>

Digitalisation impact

21/04/2023

Contents

[Introduction 3](#_Toc135581248)

[Exploring the data 4](#_Toc135581249)

[Scatter plots 4](#_Toc135581250)

[Histogram 5](#_Toc135581251)

[Distributions 5](#_Toc135581252)

[Conclusion 6](#_Toc135581253)

[Unsupervised learning 6](#_Toc135581254)

[Hyperparameters 6](#_Toc135581255)

[K-means 7](#_Toc135581256)

[Hierarchical clustering 9](#_Toc135581257)

[Conclusions 9](#_Toc135581258)

[Supervised learning 10](#_Toc135581259)

[Essence 10](#_Toc135581260)

[Hyperparameters 10](#_Toc135581261)

[Confusion matrix 11](#_Toc135581262)

[ROC-analysis 12](#_Toc135581263)

[Conclusion 12](#_Toc135581264)

[Annex 13](#_Toc135581265)

[References 14](#_Toc135581266)

# Introduction

The dataset used in this practical assignment has been obtained via the following link: [UCI Machine Learning Repository: Algerian Forest Fires Dataset Data Set](https://archive.ics.uci.edu/ml/datasets/Algerian+Forest+Fires+Dataset++)

The authors, Faroudja Abid & Nouma Izeboudjen, have developed an "Algerian Forest Fires Dataset Data Set" dataset. The problem domain of this dataset includes forest fire prediction and fire management strategies. Furthermore, there is no information about the licensing aspects of the dataset. The dataset was obtained: "We have used the meteorological observations for the summer of 2012, from June to September, since the ﬁre occurrence is high in this period and 2012 is the year where the recorded ﬁre occurrence is the highest from 2007 to 2018." (F. Abid & N. Izeboudjen, 2020). The dataset contains 243 instances and 14 features. The features are divided into two roles, namely numerical (day, month, year, temperature, rain, RH, Ws, FFMC, DMC, DC, ISI, BUI and FWI) and categorical (classes) see Table 1. The classes have two values: no fire given 0 (106 instances) and fire given 1 (137 instances) (see Figure 1).

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Meaning | Value type | Range of values |
| Day | Day of the observation | Numerical | 01 to 31 |
| Month | Month of the observation | Numerical | 06 to 09 |
| Year | Year of the observation | Numerical | 2012 |
| Temperature | Temperature noon | Numerical | 22 to 42 |
| Rain | Rain of the day in mm | Numerical | 0 to 16.8 |
| RH | Relative humidity in % | Numerical | 21 to 90 |
| Ws | Wind speed in km/h | Numerical | 6 to 29 |
| FFMC | Fine fuel moisture code index from the FWI system | Numerical | 28.6 to 92.5 |
| DMC | Duff moisture code index from the FWI system | Numerical | 1.1 to 65.9 |
| DC | Drought code index from the FWI system | Numerical | 7 to 220.4 |
| ISI | Initial spread index from the FWI system | Numerical | 0 to 18.5 |
| BUI | Buildup index from the FWI system | Numerical | 1.1 to 68 |
| FWI | Fire weather index | Numerical | 0 to 31.1 |
| Classes | Fire or no fire | Categorical | 0 or 1 |

Table 1: Features with explanations.

A picture containing text, screenshot, number

Description automatically generatedA picture containing text, screenshot

Description automatically generated

Figure 1: Snippet of the structure of the data

# Exploring the data

## A graph with red and blue dots Description automatically generated with low confidenceScatter plots

A picture containing text, screenshot, plot, line

Description automatically generatedFigure 2: FFMC – ISI scatter plot Figure 3: FWI – ISI scatter plot

The first scatter plot I decided to analyse is the fine fuel moisture code and initial spread index. The graph shows an exponential relationship between the two variables see figure 2. Furthermore, class separability is clearly seen. The no-fire incidents have a quasi-linear relationship with a slightly positive slope. The fire incidents clearly have a steeper slope. This shows a relatively stronger association between the two variables. The second scatterplot (see figure 2) has a linear relationship. The graph is interesting because of the class distribution. The non-fire events are clustered together in the lower left corner, while the fire incidents mostly have higher values.

## Histogram





Figure 4: DC feature statistics Figure 5: DMC feature statistics

Analysing the distribution of figure 4 and figure 5 shows a pretty similar graph. However, quite some differences are observed if we go more in detail. Both show signs of skewness with a tail to the right. This is visible on the distribution graph but also by looking at the difference between the mean and median. Figure 4 shows a more extreme case as the mode is extremely low compared to the minimum and maximum, as shown by the dispersion. Furthermore, we can see that in both figure 4 and figure 5, no fire reached a peak on the first bar and is heavily skewed right afterwards, but the value fire is more prevalent on the second bar and follows a more normal distribution.

## Distributions

A picture containing text, plot, diagram, line

Description automatically generatedA graph with red and blue lines

Description automatically generated with low confidence

Figure 6: FFMC distribution Figure 7: BUI distribution

In both distributions (see figure 6 and figure 7), we can clearly see that the classes, fire and no fire, are pretty separable. In figure 6, only a tiny fraction of around 80 FFMC overlap. As for figure 7, the overlapping is clearly more present as we can still see some blue bars after 30 BUI, but as our goal partially is to be able to predict when there is a chance that fire might occur, it is still notable to discuss the fact that when BUI is lower than 5 no fire occurred. Also, we could analyse the outliers to have a better understanding of why in these cases, no fire occurred as a way to find a method to prevent the fire. Furthermore, the spread in both figures is interesting to analyse as in figure 6, the spread is limited, and we could conclude we need a high FFMC. On the other hand, in figure 7, contradictory, the spread is much more significant, as we can see with the standard distribution line, with only the conclusion that a BUI lower than 5 is safe and that a high BUI is rarer.

## Conclusion

Both classes, fire and not fire, are pretty well balanced as 44% of the data is no fire, and 56% is fire cases. Furthermore, we have seen that for all features, classes are overlapping. However, this is only slightly the case for some features, which will help us with the separability of the classes. For example, if further analysis was done, we might discover a third class with inter-signs. This class would be an observation class where both events could occur. However, in this case, the classes are pretty close to each other and might be challenging to separate. Overall, the non-fire events seem to be more grouped with some outliers as the fire events seem to have a broader range. In conclusion, the features FFMC, DC, BUI, DMC, ISI, FWI, and the classes will be considered for further analysis.

# Unsupervised learning

## Hyperparameters

The number of clusters, distance metric, scaling or normalisation, linkage method and pre-processing steps are the hyperparameters that will impact our studies. First, the number of clusters will be chosen using the k-means and the correlation of the features. By doing a trial-and-error analysis of the features to include as to reach a higher silhouette score, it represents a better fit for the number of clusters. The number of clusters affects the interpretability, the computational complexity, the separation and granularity of the data, the homogeneity or heterogeneity within the cluster, and the risk of over- or underfitting. Secondly, the distance metric affects how the data is sorted, compared, and evaluated since it quantifies the similarity or dissimilarity between data points. This hyperparameter will be essential for hierarchical clustering. Thirdly, scaling or normalisation, in this case, both methods we will discuss later, use normalisation. Its effects make it easier to compare the data while not affecting the interpretability. The algorithm's performance can also be improved, as well as the data visualisation. Fourthly, the linkage method used for the hierarchical clustering algorithm helps measure the dissimilarity between clusters. The choice of method will affect the cluster shape ad size, the sensitivity to outliers, the way the clusters are merged, the cohesion and interpretability, the computational complexity, and the balance between local and global similarities. Lastly, although normalisation is also part of pre-processing, it includes a lot more. The pre-processing has a significant effect on the data quality and consistency, feature engineering (not applicable in this project), noise and dimensionality reduction, and imbalance of the data.

## K-means

A screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidence

Figure 8: k-means 6 features Figure 9: k-means 5 features

As mentioned previously, when exploring the data, we mentioned further analysis of 6 features and the classes. Nevertheless, as we can deduce from the silhouette scores in figure 8 and figure 9, we decided to reduce the features to 5 as it proved to augment the score by 0.08. The feature that has been taken out of consideration is the FFMC. The silhouette score ranges from -1 to 1, but the closer to 1, the better matched the data point is to its own cluster. Unsurprisingly, we will continue with a number of clusters of 2 as it maximises the score. An average score of 0.574 indicates that the data points are reasonably well grouped and are more closely spaced from one another within their cluster than from points in other clusters. There may still be some overlap or ambiguity in the grouping, but it suggests a decent amount of cohesion within clusters and some distinction between clusters. This will be seen further in the analysis.

A picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generatedFigure 10 (left): Silhouette plot based on clusters for 6 features

Figure 11 (right):

Silhouette plot based on clusters for 5 features

A picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 12 (left): Silhouette plot based on classes for 6 features

Figure 13 (right): Silhouette plot based on classes for 5 features

A picture containing text, screenshot, colorfulness

Description automatically generatedA picture containing text, screenshot, colorfulness

Description automatically generatedAs you can see, figures 10 to 13 show the different silhouette plots obtained. For the silhouette plot based on clusters, the plot with 5 features holds more precision than the one with 6 features. Despite that, the silhouette plots based on the classes have awful results. Over half of the fire events are misclassified. In figure 12 more fire events are classified correctly compared to figure 13, which has more non-fire events organised correctly. However, as the proper classification of non-fire events is still low in figure 12 we decided that figure 13 holds more power.

A picture containing text, screenshot, colorfulness

Description automatically generatedFigure 14: Scatter plot BUI-FWI based on silhouette score

Figure 15: Scatter plot BUI-FWI based on classes

Figure 16: Scatter plot BUI-FWI based on clusters

Figures 14 to 16 are closely related. We can see that the region in figure 14 with a lower silhouette score is the most different region between figures 15 and 16. This also means that this area holds the most errors in the clustering. The only exception is the upper part of the yellow zone of the scatterplot, which was misjudged as the conditions are close to non-fire events. This is a zone that represents well the third class we suggested earlier.

## A picture containing text, screenshot, diagram, display Description automatically generatedA picture containing text, screenshot, software, graphics software Description automatically generatedA picture containing text, screenshot, display, software Description automatically generatedHierarchical clustering

A picture containing text, screenshot, diagram, software

Description automatically generatedFigure 17 (left upper corner): hierarchical clustering of 6 features 1.6%

Figure 18 (right upper corner): hierarchical clustering of 5 features 2.6%

Figure 19 (left under corner): hierarchical clustering of 6 features 20.1%

Figure 20(right under corner): hierarchical clustering of 5 features 20.6%

First, the different linkage methods were considered and tried for hierarchical clustering. It has been decided to continue with the weighted linkage method, which holds the best results. In figures 17 to 20 we can see the clustering of 6 features (left) and 5 features (right). As we can see, the results do not hold any significant results. The upper figures show the height ratio when no class has been mis-clustered. As we can see, 1.6 and 2.6% represent low dissimilarity levels. The figures underneath represent satisfying height ratios with an average number of mis-clustering, but the difference between both is extremely low, with only 0.5% difference.

## Conclusions

In conclusion, a lot of work is needed to identify the class based on data correctly. Figure 14 holds the most power as it clearly shows a better understanding of the region that needs to be analysed further or even differentiated. As we all know, fire is a natural occurrence, and these occurrences can always surprise us even if researched well. However, We should be aware that human factors can impact the results and may partially be the reason for this third class we are suggesting.

# Supervised learning

## Essence

This analysis will use three types of supervised learning algorithms: kNN, random forest and neural networks. K-Nearest Neighbors (kNN) is a straightforward supervised learning technique that predicts the class or label of a new data point based on the majority vote of its k nearest neighbours. It considers similarities between occurrences and bases predictions on the immediate vicinity of the data points. The selection of k influences the flexibility of the decision boundary. kNN is instance-based, non-parametric, and makes no assumptions about the distribution of the underlying data. We chose it because it can handle classification and regression problems and is clear to understand. Random Forest is an ensemble learning system that uses numerous decision trees to produce predictions. It creates several decision trees by bagging and randomly choosing features. Each tree is trained using a separate bootstrap sample and a subset of characteristics to add variety and lessen overfitting. For classification or regression problems, Random Forest utilises the majority vote or averaging of individual tree forecasts. It is renowned for its dependability, capacity for managing large-scale data, and feature importance estimation. Random Forest was chosen for its ability to handle complicated issues while avoiding overfitting and delivering insights about feature relevance. Using both techniques, we can investigate various supervised learning strategies. kNN offers a simple and understandable approach, whereas Random Forest uses the strength of ensemble learning. In terms of ease of use, adaptability, and performance, these methods complement and offer a more comprehensive approach to handling supervised learning problems.

## Hyperparameters

A screenshot of a computer

Description automatically generated with medium confidenceA picture containing text, font, screenshot, number

Description automatically generatedThe hyperparameters of the supervised learning algorithms are more specific to the algorithms see table 2. The kNN's hyperparameter is k or the number of neighbours considered. In our case, k is equal to 5. The random forest has different parameters: n estimators, max features, max depth and min samples split. N estimators are 10 number of trees; in this case, the smallest sample split is 5. However, we did not set a maximum depth or a maximum amount of features—lastly, the neural network. We decided to go with 100 neurons in the hidden layers for the neural network and to use the rectified linear unit activation function as it held better results than the hyperbolic tangent. Furthermore, the regularisation is set at 0.0001 and the maximum number of iterations at 200. We also checked that the algorithms would be replicable. These parameters help optimise the accuracy, complexity, fitting, and generalisation of the model.

Figure 21 (left): test on data-test

Figure 22 (right): cross-validation test

We can see in figure 21 that the three-learning algorithm shows excellent results with first place winner in the random forest. However, in figure 22 we can see that the cross-usage of the kNN and the neural network algorithm holds the best cross-validation results. These two results explain our choice of algorithms. Lastly, we decided to add 30% of our data to the test sample or 72 instances (31 non-fire events and 41 fire events or, respectively 43 and 57%). The training data set held 70% or 171 instances (75 non-fire events and 96 fire events, respectively 44 and 56%).

|  |  |  |
| --- | --- | --- |
| **Learning algorithm** | **Hyperparameter** | **Value** |
| kNN | k/number of neighbour | 5 |
| Random forest | N estimator | 10 |
|  | Smallest sample split | 5 |
| Neural network | Neurons in hidden layer | 100 |
|  | Activation function | ReLu |
|  | Regularisation | 0.0001 |
|  | Max number of iteration | 200 |

Table 2: Hyperparameters applied

## Confusion matrix

A screenshot of a graph

Description automatically generated with low confidenceA screenshot of a graph

Description automatically generated with low confidenceA screenshot of a graph

Description automatically generated with low confidence

Figure 23 (left): confusion matrix for kNN

Figure 24 (middle): confusion matrix for random forest

Figure 25 (right): confusion matrix for neural network

Figures 23 to 25 show that the random forest and neural network have the same results. These results are slightly better than the kNN's results as we did not predict any fire event that did not happen. However, in all cases, mistakes happen where non-fire events were predicted when an actual fire occurred. This could still hold a positive value as we would be able to prepare adequately for all the predicted cases. Nonetheless, it would be interesting to analyse the mistakes more in detail and see if they are from the suggested third-class type.

## A screen shot of a graph Description automatically generated with medium confidenceROC-analysis

Figure 26: ROC-analysis

Figure 26 shows the ROC analysis offers the best results for the random forest algorithm followed by the neural network and the kNN. All models show high marks for the analysis; however, we should be aware that this analysis focuses on performance across all values, not specific requirements. Furthermore, the ROC analysis assumes that FP and FN hold equal values, which is not the case, as discussed in the confusion matrix section.

## Conclusion

Overall, the supervised learning algorithm seems to have good results compared to the unsupervised algorithms in both models. Using supervised algorithms, we used the labelled data to direct the learning process and produce accurate predictions about unobserved cases. Using this strategy, we could accurately classify the data and create forecasts by using its built-in relationships and patterns.

# Annex

A picture containing text, diagram, map

Description automatically generated

Figure 27: overview of Orange

# References

Abid, F., & Izeboudjen, N. (2019). *Algerian Forest Fires Dataset Data Set*. UCI Machine Learning Repository: Algerian Forest Fires Dataset Data set. [UCI Machine Learning Repository: Algerian Forest Fires Dataset Data Set](https://archive.ics.uci.edu/ml/datasets/Algerian+Forest+Fires+Dataset++)

Abid, F., & Izeboudjen, N. (2020). *Predicting forest fire in Algeria using data mining techniques: Case Study of the Decision Tree Algorithm*. ResearchGate. <https://www.researchgate.net/publication/339062373_Predicting_Forest_Fire_in_Algeria_Using_Data_Mining_Techniques_Case_Study_of_the_Decision_Tree_Algorithm>

Gross, J. L., & Yellen, J. (1999). *Graph theory and its applications*. CRC.

Hastie, T., Friedman, J., & Tisbshirani, R. (2017). *The elements of Statistical Learning: Data Mining, Inference, and prediction*. Springer.

Leskovec, J., Rajaraman, A., Ullman, J. D., Wang,Bin, & Wang,Dakan. (2021). *Si Tan Fu Shu Ju Wa Jue Jiao Cheng = mining of massive datasets*. Ren min you dian chu ban she.