STATS 4M03/6M03: Multivariate Analysis

Final Project

Early Detection of Forest Fires in Algeria Through Statistical Analysis

Submitted to

Dr. Eman M.S. Alamer

Department of Mathematics and Statistics

McMaster University

Hamilton, Ontario, Canada L8S 4K1

November 22, 2024

Reported by

Samantha Yang (400372854)

Junbo Tan (400310355)

Matthew Li (400314681)

Yoonah Kim (400410242)

Yunxin Li (400323756)

1 Introduction

1.1 Abstract

Countries located in the Mediterranean Basin, such as Algeria, Portugal, and Spain, have a high susceptibility to forest fires during summer months (Bento-Gonçalves, 2021). Our project aims to analyze and predict the occurrence of forest fires in Algeria using a multivariate dataset that focuses on important meteorological variables influencing the risk of these fires. Using various techniques taught in STATS 4M03, we investigate patterns and relationships to predict the likelihood of forest fires occurring under various weather conditions. The objective is to provide insights into predictive factors and assess their potential for supporting early fire detection strategies, ultimately mitigating the economic, social, and ecological impacts of forest fires in Algeria.

1.2 The Data

The dataset "Algerian Forest Fires" was sourced from the UC Irvine Machine Learning Repository. The dataset was donated on October 21, 2019 and contains 244 instances of forest fire data from two distinct regions in Algeria: Bejaia and Sidi Bel Abbes. The data collection period spans from June to September 2012, with 122 instances collected from each region. The dataset was originally fragmented into two sections for each of the regions. However, we modified the dataset to include the region as its own categorical variable with the labels "Benjaia" or "Sidi-Bel Abbes" in order to have one full dataset. The dataset utilizes 11 attributes and 1 output attribute, which describes whether a forest fire occurred or not. The target variable, "Classes", categorizes instances as either "fire" or "not fire", with 138 instances classified as "fire" and 106 as "not fire". All variables other than Classes, Region, and Date are continuous. The 11 attributes are classified as:

1. Date : DD/MM/YYYY (Split into 3 variables in the dataset)
2. Temperature: Temperature measured at noon in degrees Celsius
3. RH : Relative Humidity in %
4. Ws : Wind speed in km/h
5. Rain: Total precipitation for the day in mm

The remaining attributes are components of the Fire Weather Index (FWI); a meteorologic

index used globally to predict the risk of fires where higher FWI indicates higher chances of fires (Natural Resources Canada, n.d.).

6. Fine Fuel Moisture Code (FFMC): Moisture amount in litter and other flammable matter
7. Duff Moisture Code (DMC): Moisture amount in loose, decomposing organic matter
8. Drought Code (DC): Moisture amount in compact organic material below earth's surface
9. Initial Spread Index (ISI): A fire's rate of spread based on Ws and FFMC
10. Buildup Index (BUI): Amount of fuel available for combustion based on DMC and DC
11. Fire Weather Index (FWI): How intense a fire is, based on ISI and BUI
12. Classes: Output attribute, characterized by "fire" or "not fire"

1.2.1 Exploratory Data Analysis (EDA)

To begin, we note that the UC Irvine Machine Learning Repository stated the dataset does not contain any missing values. This was confirmed as the sum of missing values in R was zero.

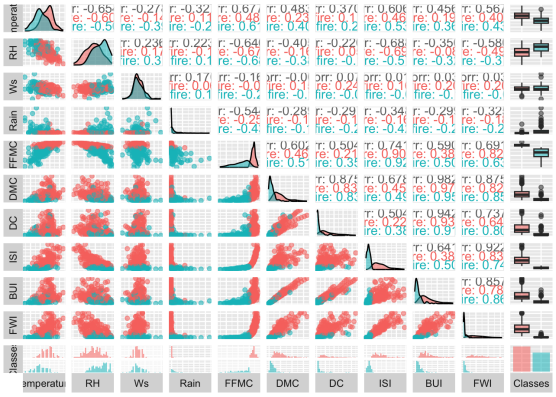

Figure 1: Pairs Plot of All Variables in Dataset

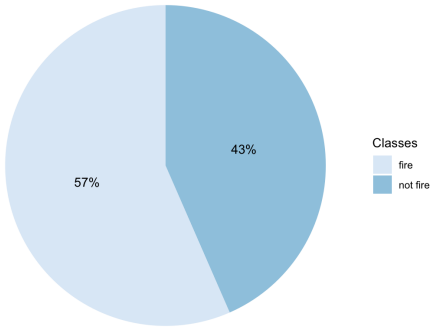

Figure 2: Pie Chart: Proportion of "Fire" and "Not Fire" From Data

It is evident that many of our variables are highly correlated such as the relationships between DMC and BUI or between DC and BUI. Both positive and negative correlations are demonstrated in our pair plot which offers the most insight on the behaviour of these variables compared to other plots. The number of instances classified as "fire" or "not fire" was also confirmed by our pie chart where we see a slightly greater proportion being classified as "fire". Based on the definition of our variables, we can see the strong correlation between certain variables is to be expected and we can

use these correlations as a baseline when determining the influence of variables on each other and on our ability to predict forest fires.

1.2.2 Data Preparation

During data preparation, categorical variables such as "Region", and irrelevant information such as "Date", were removed to not disrupt certain analyses in R. The variable "Classes" was used as the label for the data and any white spaces were removed. As there were no missing values, both imputation and removal were not necessary. However, one of the rows in the dataset was not entered correctly with multiple observations in a single cell in the .csv file, so that was fixed. There were also no extremely significant outliers to note. Lastly, when splitting our data for supervised learning, all analyses followed a 70%/30% split for training and testing, respectively. All analyses were conducted using R version 4.4.2 (2024-10-31).

2 Methodology

2.1 Hierarchical Clustering

Hierarchical Clustering is a method of unsupervised learning used to group similar objects into clusters (Alamer, 2024). It forms a “tree-like” dendrogram and shows how much the variables are separated or merged. Four different hierarchical clustering methods were observed and based on the cluster dendrograms and the Adjusted Rand Index (ARI) output, the complete linkage and Ward.D2 methods were selected.

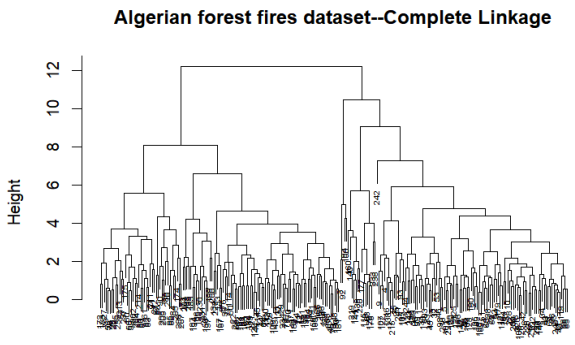

Figure 3: Complete Linkage

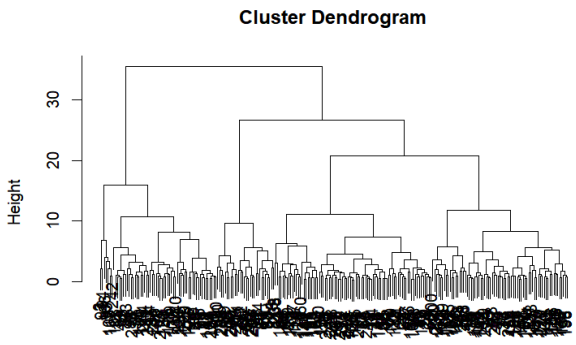

Figure 4: Ward.D2 Linkage

Based on cluster shape, interpretability, and the ARI values, we determined that the best model was the Ward.D2 linkage, followed by the complete linkage. The dendrogram for Ward.D2 was cleaner than that of complete linkage, with less chaining visible. Their respective ARI values were 0.3939026 and 0.280948.

2.2 Centroid Clustering Methods - k-Medoids

In unsupervised machine learning, k-medoids is a partitional centroid-based clustering method that can be used for data clustering. It forms clusters by using Euclidean dissimilarity, which minimizes the distance between data points within each cluster. Unlike k-means, k-medoids chooses actual data points as representatives for each of the cluster centers (Alamer, 2024).

Similar to k-means, the k-value in k-medoids represents the number of clusters the algorithm will create. In this project, the optimal k-value was determined by running the k-medoids algorithm with different values of k, ranging from 2 to 5 (Alamer, 2024). Using the silhouette method (Figure 5), k=2 was identified as the best choice, as it achieved the highest average silhouette width. The ARI for k-medoids is 0.530272, indicating moderate accuracy in the clustering process.

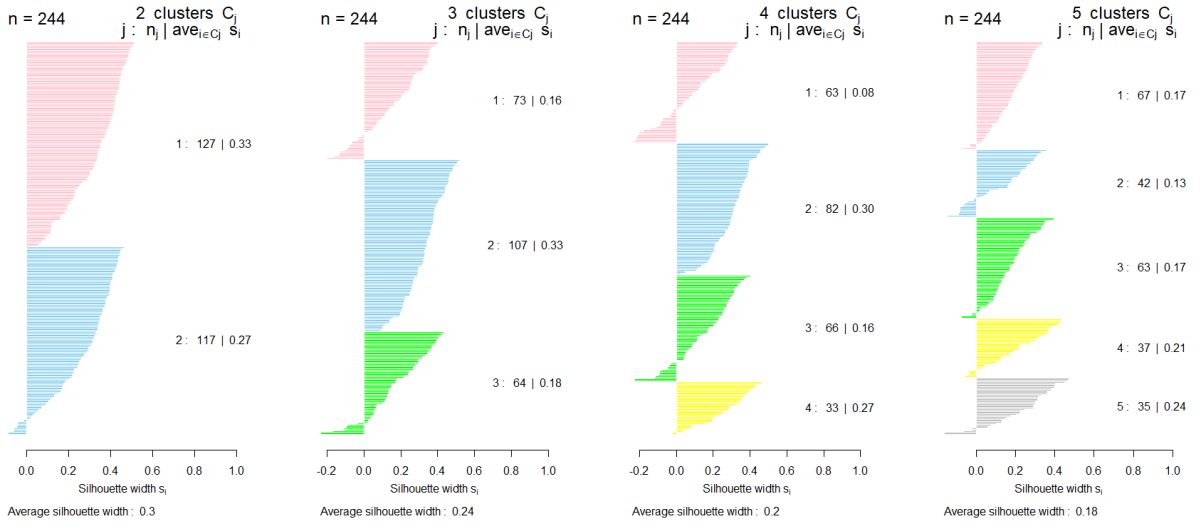

Figure 5: Silhouette width of K-medoids from k=2 to 5

2.3 Logistic Regression

Logistic Regression can be introduced as a supervised learning method as the response variable is binary in our dataset (Alamer, 2024). For the convenience of calculation and interpretation, it is worth parameterizing the response variable Class to 1 ("Fire") and 0 ("Not Fire"). This allows us to compare the full model from Table 1 and the fitted model from Table 2.

	Estimate	Std. Err.	z value	Pr(>|z|)
(Intercept)	2.657e+01	5.296e+04	0.001	1
Temperature	2.187e-09	4.818e+04	0.000	1
RH	-5.410e-09	4.834e+04	0.000	1
Ws	-7.828e-10	3.180e+04	0.000	1
Rain	4.926e-10	3.126e+04	0.000	1
FFMC	-8.302e-09	6.273e+04	0.000	1
DMC	-7.215e-09	2.095e+05	0.000	1
DC	1.267e-08	1.208e+05	0.000	1
ISI	3.839e-08	6.280e+04	0.000	1
BUI	-1.824e-08	2.982e+05	0.000	1
Classesnot fire	-5.313e+01	1.037e+05	-0.001	1

Table 1: Logistic Regression Results for full model

	Estimate	Std. Err.	z	p	Odds Ratio
intercept	0.9056	0.2865	3.160	0.002	-
Temperature	0.8675	0.2396	3.620	< 0.01	2.380846
DC	2.2660	0.4714	4.807	< 0.01	9.640362
p-value=1.825222e-20					

Table 2: Logistic Regression Results for fitted model

	1	0
1	26	7
0	5	34

Table 3: Classification table

It is obvious that the full model with all features is not statistically significant from Table 1. Therefore, **forward selection** can be used for feature selection to ensure the significance of the logistic regression. The new fitted model is demonstrated from Table 2, with the expression

$$\log\left(\frac{\pi}{1-\pi}\right) = 0.9056 + 0.8675x_1 + 2.2660x_2$$

Meanwhile, the fitted model can be tested using the likelihood ratio test under a chi-square distribution with p-value = 1.825222e-20 from Table 2. Here, we would reject H_0(all β=0). The

corresponding misclassification rate (MCR) can be calculated from Table 3, which is 0.1666667. To quantify the effect of the given features on the outcome, we use the odds ratio. The odds ratio of Temperature is 2.380846, indicating that a forest fire is 2.380846 times more likely to happen as Temperature increases by one degree Celsius. The odds ratio of DC is 9.640362, indicating that a forest fire is 9.640362 times more likely to happen as DC (Drought Code) increases by one unit.

2.4 Discriminant Analysis (LDA vs. QDA)

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) were compared as two supervised learning methods. LDA draws linear boundaries between groups in the data, whereas QDA draws boundaries in a quadratic manner, and these boundaries are based on the probability of an observation being in the class (Alamer, 2024). These methods used the full model, including all variables in the data. Tables 4 and 5 are the classification tables for LDA and QDA, respectively. Figures 6 and 7 are the partition plots for LDA and QDA.

	fire	not fire
fire	38	3
not fire	3	28

Table 4: Classification table for LDA, full model.

	fire	not fire
fire	41	0
not fire	3	28

Table 5: Classification table for QDA, full model.

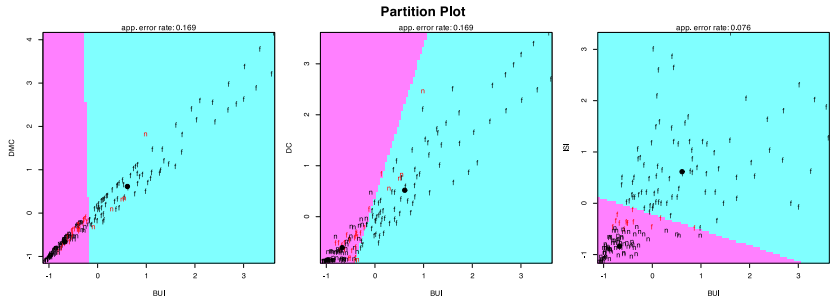

Figure 6: Partition plots for LDA with variables DMC and BUI.

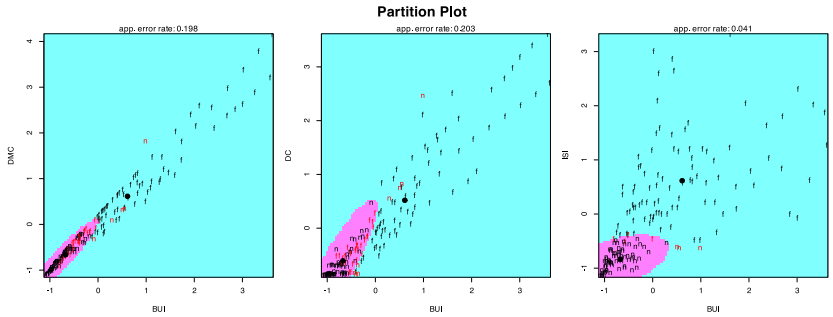

Figure 7: Partition plots for QDA with variables DMC and BUI.

Next, LDA and QDA were conducted on a reduced model: `Class~Temperature+DC`, and these variables were chosen based on the result of forward selection from Section 2.3. The following classification tables resulted:

	fire	not fire
fire	34	7
not fire	5	26

Table 6: Classification table for LDA, reduced model.

	fire	not fire
fire	30	11
not fire	7	24

Table 7: Classification table for QDA, reduced model.

3 Discussion

The ARI was used to compare clustering results as label-switching issues in clustering can negatively affect the misclassification rate (MCR). In hierarchical clustering, the best model was Ward.D2 linkage, as its ARI value was higher than that of complete linkage. However, the comparison between Ward.D2 (ARI = 0.3939026) and k-medoids (ARI = 0.530272) suggested that k-medoids was the optimal clustering method used in this project.

Logistic regression predicted the result using the features Temperature and DC as the full model for logistic regression was not statistically significant for every feature when looking at the p-values. After the prediction of fitted logistic regression using test set, it demonstrated an MCR of

0.16667. Next, we compare LDA and QDA using the MCR. For the full model, LDA had an MCR of 0.08333 while QDA had an MCR of 0.04167. If using the full model, we can conclude that QDA was better, although both methods performed well. Using the reduced model, LDA had an MCR of 0.16667 while QDA had an MCR of 0.25, lower than the full model. This LDA produced an equal MCR to logistic regression; these methods performed equally well. With less variables, LDA performed better. This is because it is easier to slice the data in a linear fashion when there are less variables. When many continuous variables exist in the full model, QDA works better to fit the data.

4 Conclusion

For the Algerian Forest Fires dataset, four machine learning methods were used, including two supervised (hierarchical clustering and centroid clustering) and two unsupervised (discriminant analysis and logistic regression). Based on the ARI values from hierarchical and centroid clustering, we conclude that k-medoids performed better. Using all variables for LDA and QDA, QDA performed better. Using a reduced model, LDA performed better than QDA. Logistic regression, using a reduced model, performed equally as well as LDA.

Using our machine learning methods, we are now able to predict the likelihood of future forest fires based on the covariates found in the dataset. In particular, models derived from classification can be used to extrapolate to new data. For future work, we could explore other clustering methods that achieve an ARI value greater than 0.6 as a higher ARI value indicates better clustering performance. The classification methods discussed in this report demonstrate good performance, but there is still room to explore other approaches for potential improvement.

5 Bibliography

Abid, F. (2019, October 21). *Algerian Forest Fires*. UC Irvine Machine Learning Repository. https://archive.ics.uci.edu/dataset/547/algerian+forest+fires+dataset

Alamer, E. (2024, September 27). *STATS 4M03/6M03: Multivariate Analysis. Lecture 8: Partitioning Methods I - Centroid Based Clustering*. [PowerPoint Slides]. Avenue to Learn.

https://avenue.cllmcmaster.ca

Alamer, E. (2024, October 2). *STATS 4M03/6M03: Multivariate Analysis. Lecture 9: Comparing Partitions*. [PowerPoint Slides]. Avenue to Learn. https://avenue.cllmcmaster.ca

Alamer, E. (2024, October 25). *STATS 4M03/6M03: Multivariate Analysis. Lecture 13: Discriminant Analysis*. [PowerPoint Slides]. Avenue to Learn. https://avenue.cllmcmaster.ca

Alamer, E. (2024, November 8). *STATS 4M03/6M03: Multivariate Analysis. Lecture 17: Binary Logistic Regression*. [PowerPoint Slides]. Avenue to Learn. https://avenue.cllmcmaster.ca

Alamer, E. (2024, November 15). *STATS 4M03/6M03: Multivariate Analysis. Lecture 19: Logistic Regression: Removing and Collapsing Predictor Variables*. [PowerPoint Slides]. Avenue to Learn. https://avenue.cllmcmaster.ca

Bento-Gonçalves, A. (2021, September 1). *Algeria suffers from devastating wildfires, but faces big challenges in addressing them*. The Conversation. https://theconversation.com/algeria suffers-from-devastating-wildfires-but-faces-big-challenges-in-addressing-them-166944

Canada, N. R. (n.d.). *Canadian Wildland Fire Information System: Canadian Forest Fire Weather Index (FWI) System*. https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.