Tutorial 11 Worksheet AY 22/23 Sem 1

DSA2101

Forest fires dataset

The forest fires dataset `forestfires.csv` comes from the UCI machine learning repository. It contains information about the damage caused by forest fires in a natural park (Montesinho) in Portugal. It is of interest to be able to predict the damage (in terms of area) that a fire will cause based on meteorological data.

The columns in the data are:

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

The FFMC, DC and ISI are indicators of the possibility of a damaging forest fire ensuing; they are derived from weather data from the previous day. Hence, these columns are suitable for use in predicting the damage that a forest fire could cause.

The output variable (`area`) is highly skewed. Read the data into R, and transform `area` to `lg_area` using

$$\text{lg\_area} = \log(1 + \text{area})$$

Also create a character variable `damage`, that takes the value "damage" when `area` is positive and "no_damage" otherwise.

1. Create a visualisation using the two variables `month` and `damage`.
2. Use `GGally::ggpairs` and `Hmisc::describe` to understand more about the data. List down your observations and insights about the data.
3. Focus on the months of Mar, Aug and Sep, and create a visualisation of X, Y and `lg_area`.

Clustering Practice

Let us return to the Wisconsin Breast Cancer dataset. In this situation, there are two clusters of interest to us - the two types of diagnosis. This is a rare situation when we know the two clusters of interest to us.

Carry out hierarchical clustering and assess if the returned 2-clusters match with the labelled clusters. In a sense, this gives an indication of the value of the features in discriminating between the two diagnosis.