Name: Ahmed Aldawoud
Student ID No.: 18038024

Foundation of Information science

Assignment 2

Forest Fire Data set

1) Introduction about the data:

This data set is about fires that happen in forests. It has **10 numerical attributes**, and **two string attributes** (Day and Month the data was recorded). As for classes, it is represented by **the amount of forest area burned**.
Since area is a numerical value, I will fuzzify the data to **3 classes** which are:

1) **Small** area burned.
2) **Medium** area burned.
3) **Large** area burned.

Note: Since the purpose of my analysis is to check for how these attributes contribute to the fire spread, this means that only when the amount of area burned is not zero that the sample is relevant. So I **filtered** the data before working on it where I removed all samples where the area burned is **zero**.

To picking the right **2 attributes** for the fuzzy logic, I checked their **correlation**:

X: 0.0703
Y: 0.0502
FFMC: 0.0543
DMC: 0.0890
DC: 0.0467
ISI: 0.0021
Temp: 0.1102 (highest correlation)
Relative humidity: -0.1048 (Second highest correlation)
Wind: 0.0021
Rain: -0.0129

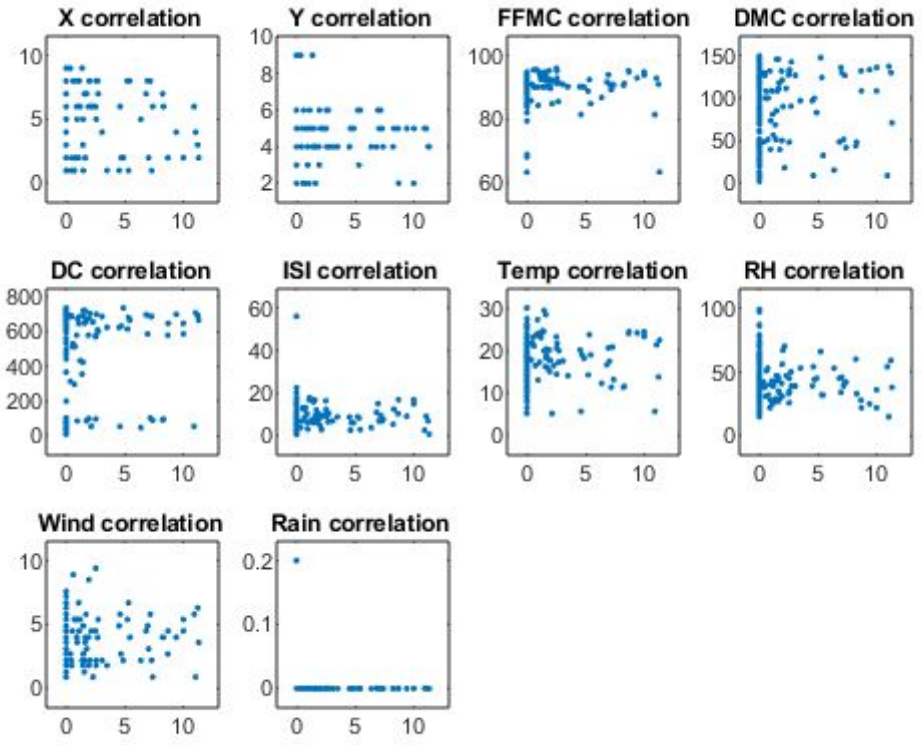
Notes about correlation:

1) Negative values doesn't mean that there's low correlation. It just means the correlation is negative.
2) Unfortunately, none of the attributes had a big correlation with the amount of area burned. This might be because most of the samples had zero area burned which makes the correlation drop this low.

Name: Ahmed Aldawoud
Student ID No.: 18038024

I picked the **most two correlated** attributes with the **Burned area** which are; **Temperature**(Positive correlation) **,** and **Relative humidity**(Negative correlation).
I makes sense to use those two because high **Temperature** would induce more fires to happen and high **Humidity** would slow down fire spread because dry forests tend to burn faster.

1.1 Correlation graph:



Note: I used 200 samples of the data picked randomly to make this plot graph. The main reason is plotting all the data would look like a mess. And plotting few samples would make the graphs look empty and might not represent most samples provided in the data.

Name: Ahmed Aldawoud
Student ID No.: 18038024

1.2 Attributes summary:

Temperature:

Temperature is measured in **Celsius degrees**.
Medium Temp.: 19.3
Maximum Temp.: 20.1
Minimum Temp.: 2.2

Relative Humidity:

Relative humidity is a **percentage** where 100% humidity means that the air is saturated with water molecules.

Medium Humidity.: 41
Maximum Humidity.: 96
Minimum Humidity.: 15

1.3 Area Burned summary:

Area is measured in **Hectare**.
Medium Area Burned.: 6.37
Maximum Area Burned.: 1090.84
Minimum Area Burned.: 0.09

Name: Ahmed Aldawoud
Student ID No.: 18038024

2) 20 Samples to make the rules:

2.1 Randomising data:

I made sure all the data is **randomised** at the beginning so every time the code runs it randomises all the data.
This means; when I take the first 20 rows of my data, they will be random.

Note: **correlation** of **rain** with **burned area** might appear as **NaN** when you run the code **sometimes**. The reason behind that is that when the samples are randomised, there's a high chance that the 20 samples that are picked have the rain as 0. Which will mess up the correlation value.

2.2 Attributes Correlation with Area Burned (for 20 samples):

X	0.1368
Y	0.0950
FFMC	0.0798
DMC	0.0756
DC	0.1238
ISI	-0.0756
Temp	0.2136
Relative Humidity	-0.2789
Wind	0.0558
Rain	NaN

Again, the most correlated attributes to the amount of area burned are;
Temperature(0.2136) and Humidity(-0.2789) (For 20 random samples). Thus, I will use them for the fuzzy logic.

Name: Ahmed Aldawoud
Student ID No.: 18038024

2.3 Summary for 20 samples:

Min Temp	5.8
Max Temp	28.2
Med Temp	21.9
Min RH	24
Max RH	82
Med RH	43
Min Area burned	0.33
Max Area burned	1090.8
Med Area burned	3.64

2.4 Fuzzifying the attributes:

Temperature:

1) Low (5.8)
2) Medium(21.9)
3) High(28.2)

Relative humidity:

1) Low(24)
2) Medium(43)
3) High(82)

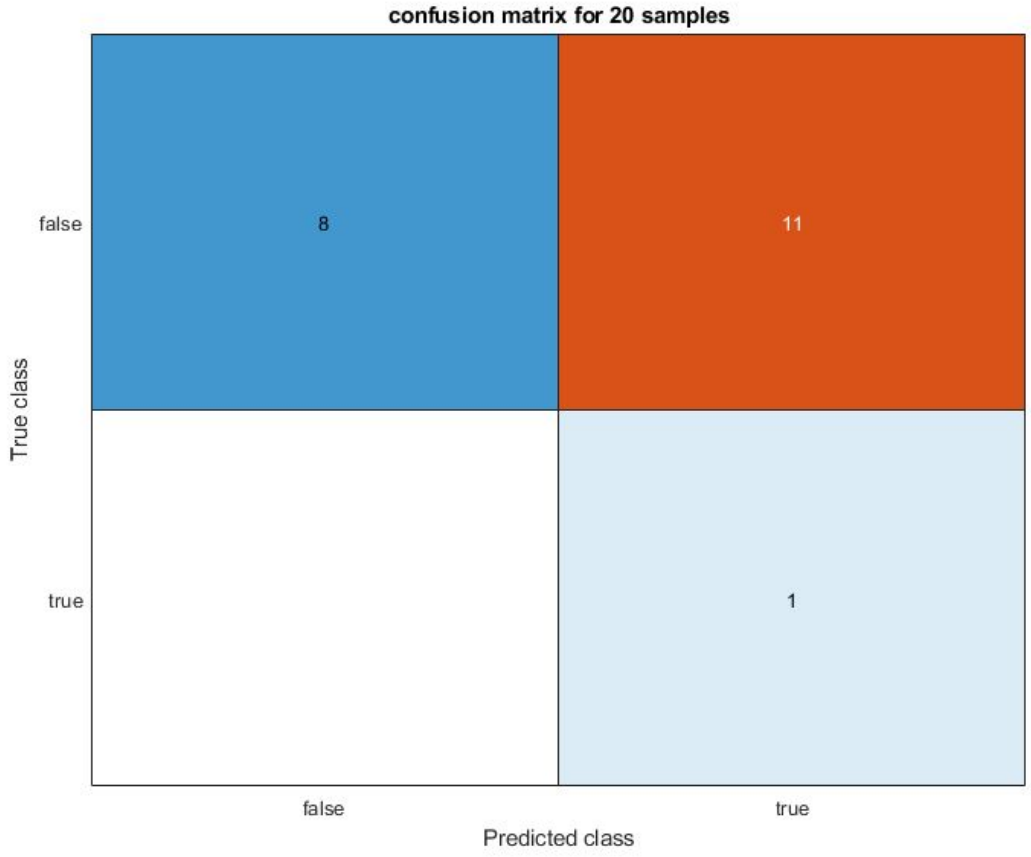
2.5 Fuzzifying Area Burned:

1) Small area burned.(3.64)
2) Large area burned.(1090.8)

Note: The reason why Burned area has 2 Fuzzified values instead of 3 is that the Median is way closer to the minimum value than to the maximum one. Which means the having a medium value is going to be really close to the small value in reality.

Name: Ahmed Aldawoud
Student ID No.: 18038024

2.6 Confusion matrix and its calculations(For 20 Samples):



Accuracy:45%

Precision:66.67%

Recall: 100%

Name: Ahmed Aldawoud
Student ID No.: 18038024

3) Testing Fuzzy logic:

3.1 Fuzzy logic notes

3.1.1 About the data correlation

Unfortunately, when I tested the Fuzzy logic I made using the 20 samples on the same 20 samples it gave a very poor accuracy of 45%. This might be because all the attributes are poorly correlated with the amount of area burned.

3.1.2 Membership functions.

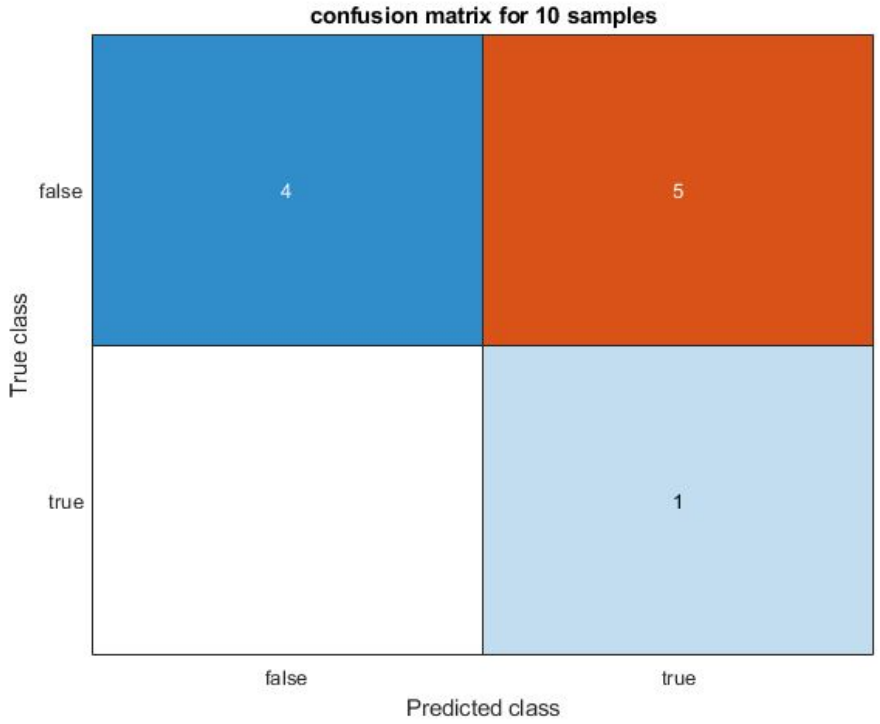
Picking the right membership function was the hardest task. I considered two factors; how many fuzzified values I had for each attribute, and how correlated they are with the amount of area burned. There are many types of membership functions and I believe there must be better ones for my data than the ones I picked, but due to the limited time I only tested few membership functions on my data.

3.1.3 Rules

The way I concluded the rules was by looking at the 20 samples and analysing how they correlate with the amount of area burned.
High humidity meant less area was burned and High Temperature meant more area was burned.

Name: Ahmed Aldawoud
Student ID No.: 18038024

3.2 Confusion matrix and its calculations(For 10 Samples):



Accuracy: 50%
Precision:66.67%
Recall:100%

Note: to read the confusion matrix:
True means **Large** area was burned. **False** means **Small** area was burned.

3.2.1 Realisations about confusion matrix

The **accuracy** is pretty low due to the low correlation of attributes.
Precision is 66.67% which means that 66% of the results were consistent.
The **Recall** is 100%. This was because no data was predicted to result in small area burned (False) when it was actually large area (True).

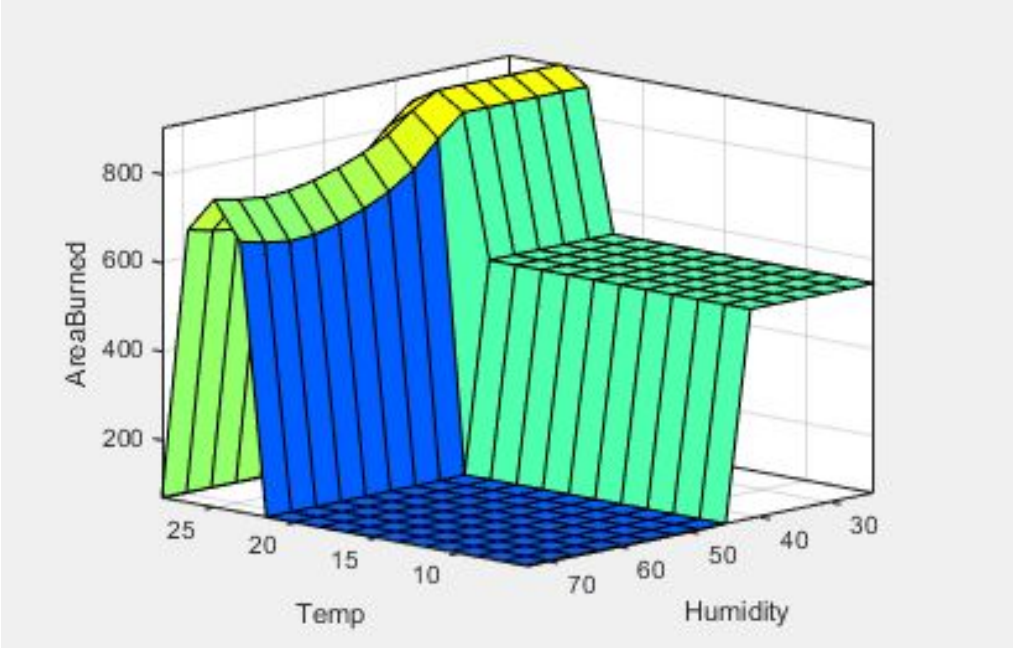
We can also notice that all the errors were when the area burned was actually small but it was predicted as large (False Positive).

Name: Ahmed Aldawoud
Student ID No.: 18038024

3.2.2 Comparison with 20 samples confusion matrix

We can notice that the results are close. There are True Negative predictions for both. The accuracy is slightly higher for the 10 samples, which might be because of randomisation, but it is still low. The precision and recall are the same for both.

3.3 Response surface



From the response surface above we can see that Humidity affects the amount of area burned negatively. This means when humidity is high area burned is small and vice versa. Temperature affects the amount of area burned positively. This means, when the temperature is high, the amount of area burned is large and vice versa.