Python for Data Analysis

Solar flares Predicition

Sebastien LIAO-DIA6

Context of the problem

- Data featuring solar flare evolution and classification
- Each row represents 1 active region of the sun
- The database contains 3 potential classes, one for the number of times a certain type of solar falre occured in a 24hrs period

Objective:

- Predict a solar flare class of a region based on the parameters given
- Use of the Zurich Sunspot Classification system

Keypoints

01 Data Presentation

Presentation of the variables in the dataset

02 Data exploration

Analysis of the data, link between the variables

03 Features engineering

04 Prediction model

Modeling prediction using scikit-learn

01

Data Presentation

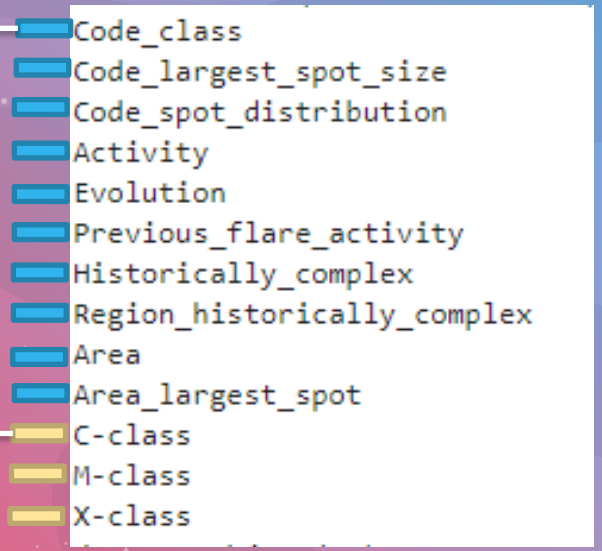
1. Data Presentation

FEATURE

Input data of the model allowing us to predict the target

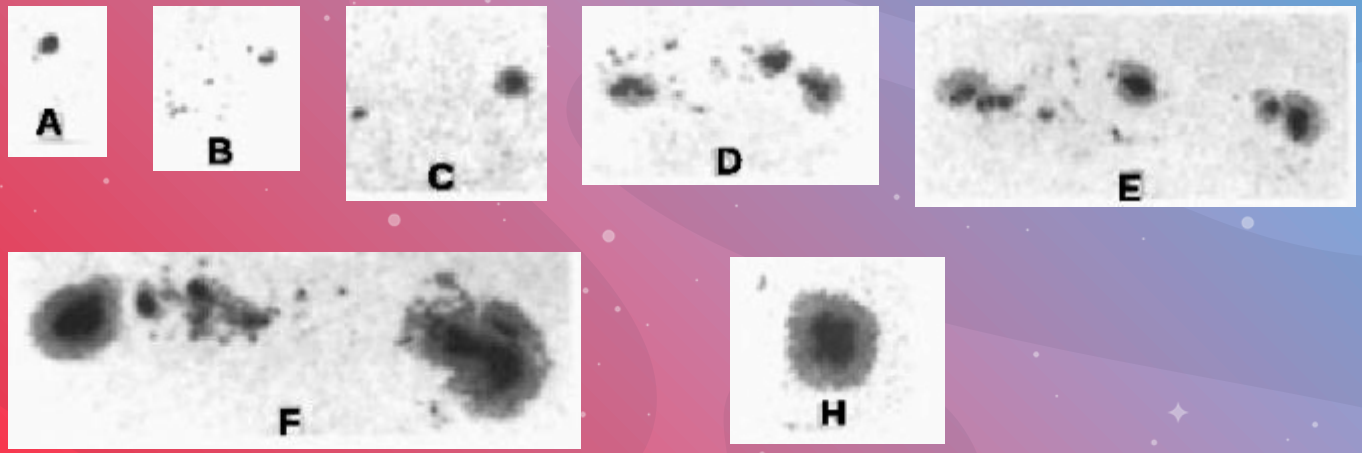
TARGET

Data we want to predict:
The number of solar flare on 1 region of the sun



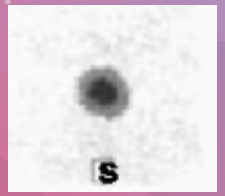
1. Data Presentation

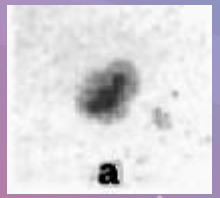
Code_class: Modified Zurich Class (A, B, C, D, E,F ,H)

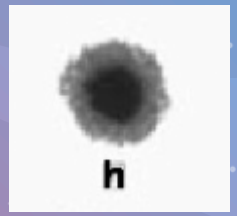


1. Data Presentation

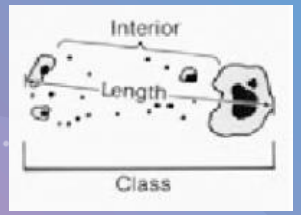
Code_largest_spot_size: Penumbra of the largest spot(X, R, S, A, H, K)

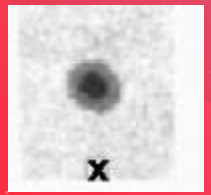


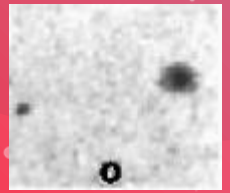


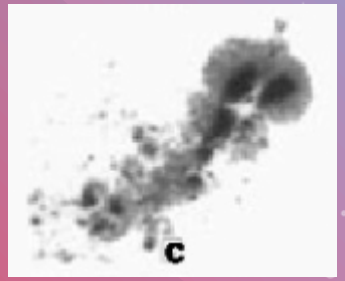


Code_spot_distribution: Sunspot distribution(X, O, I, C)









1. Data Presentation

- Activity (Reduced, Unchanged)
- Evolution (Decay, No growth, Growth)
- Previous_flare_activity (Nothing as big as an M1, One M1, More acivity than one M1)
- Historically_complex (Yes, No)
- Region_historically_complex (Yes, No)
- Area (Small, Large)
- Area_largest_spot (Less or equal than 5, More than 5)

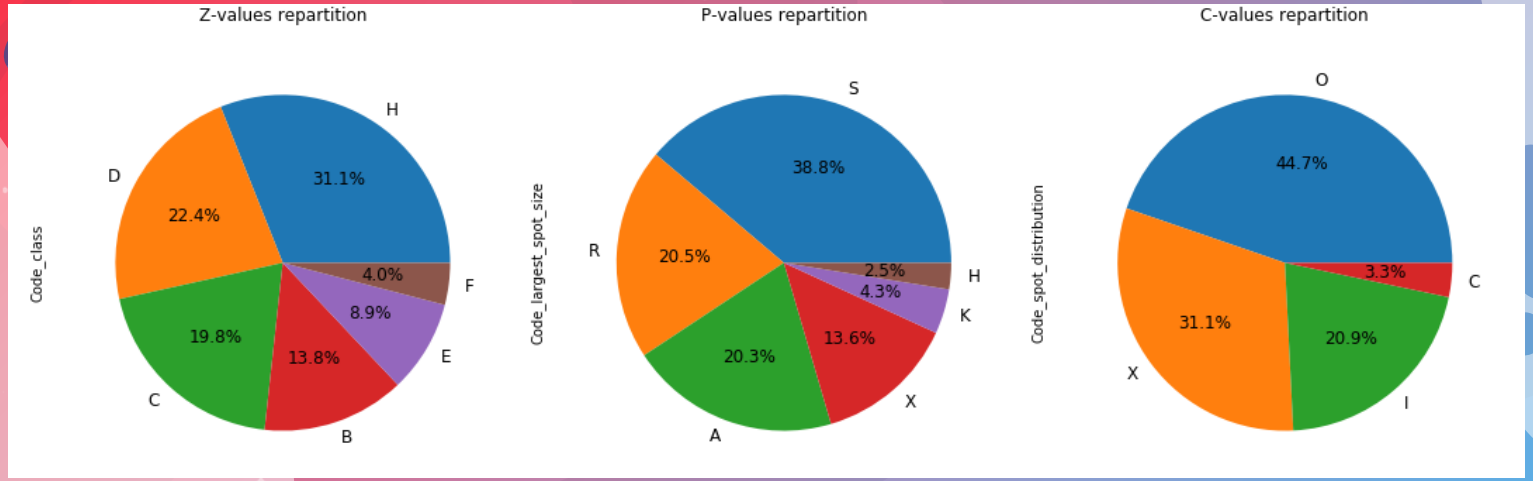
Shape: 1066 rows, 13 columns.
Features: 10 categorical variables, 3 numerical variables

02

Data Exploration

2. Data Exploration

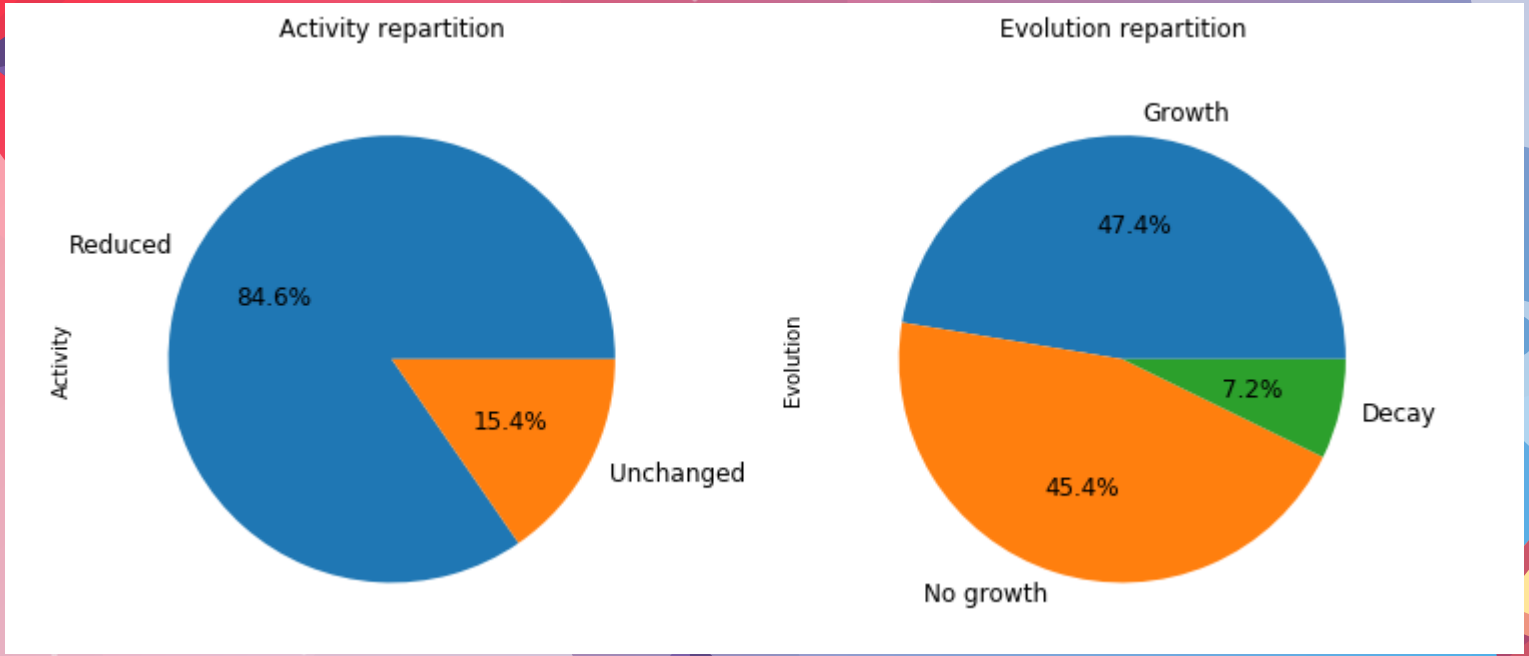
McIntosh classification repartition



We can notice that the most highest number of class for the 'zpc' form is respectively H, S and O
-H represents an unipolar sunspot group with penumbra but relatively large
-S means that the largest spot has mature, dark, filamentary penumbra of circular or elliptical shape with little irregularity to the border. It's size is rather medium
-O implies that it's one of the smallest sunspot distribution but with multiple spots

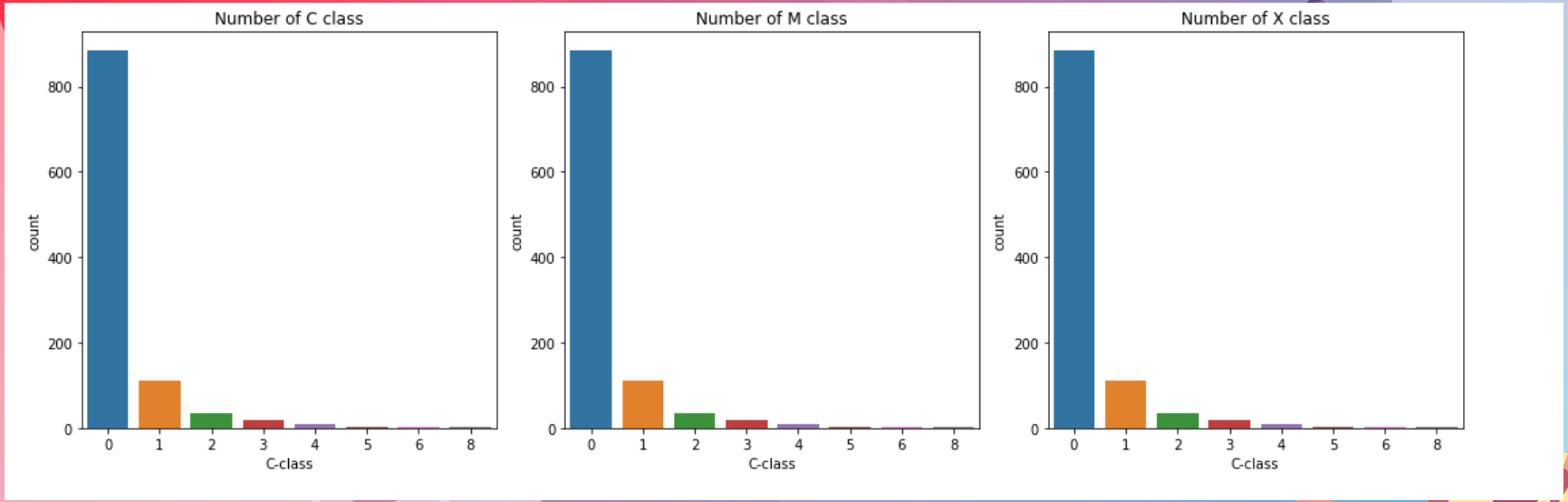
2. Data Exploration

Activity and Evolution repartition



2. Data Exploration

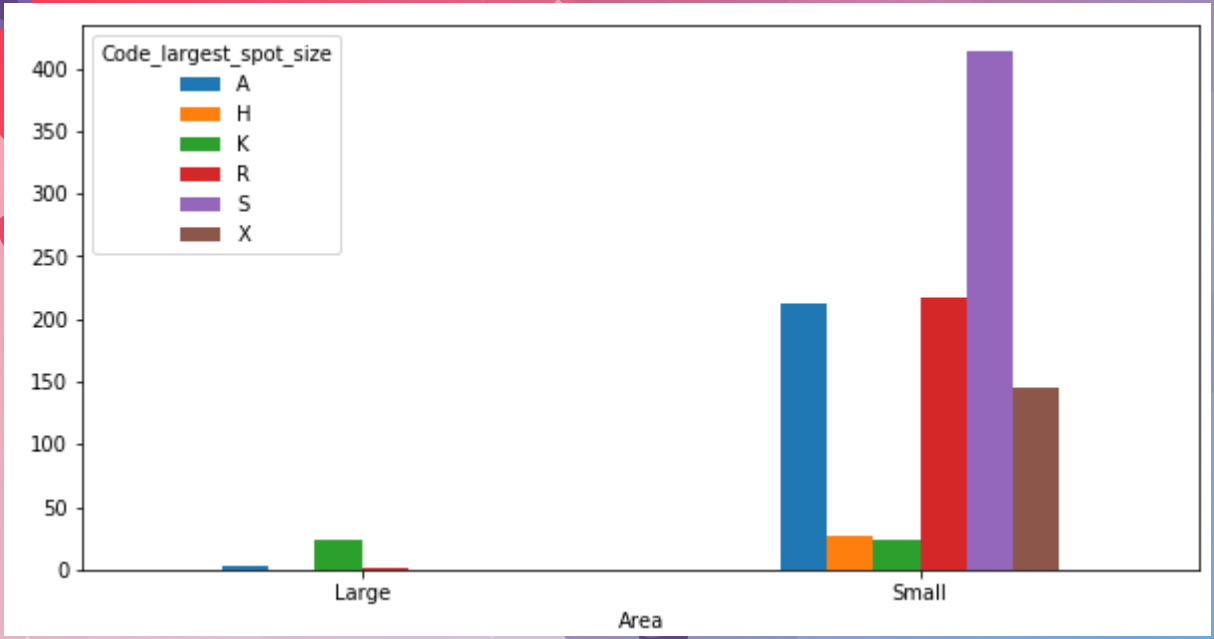
Solar flare classes count



Since there's a lot of areas where there's not a single class (Which makes sense since solar flares aren't happening all aound the sun),this could affect the prediction of the data

2. Data Exploration

Relationship between the area of the solar flare and the largest spot size



03

Features Engineering

3. Features Engineering

- Splitting the data into multiple columns
- Changing the « numerical » values into categorical values:
  - Changing the values of activity from (1, 2) to (Reduced, Unchanged)

- One Hot encoding of all the features except the number solar flares classes in order to use them in scikit learn
- Same reason for Ordinal Encoder

3. Features Engineering

- Verification of the accuracy of the model using the mean absolute error (MAE) and the mean squared error (RMSE)

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

$$MSE = \frac{\sum_{i=1}^{n} (f(x_i) - y_i)^2}{n}$$

$$RMSE = \sqrt{MSE}$$

04

Prediction Model

4. Prediction Model

Model Regression:

- Linear Regression
- Gradient Boosting Reressor
- Logistic Regression

Evaluation Function:

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

$$MSE = \frac{\sum_{i=1}^{n} (f(x_i) - y_i)^2}{n}$$

$$RMSE = \sqrt{MSE}$$

4. Prediction Model

Linear Regression:

```
First 5 LM predictions(C-class):
0  0.018472
1  0.146782
2  0.071126
3  0.417559
4  0.145373
```

```
C-class
LM MAE: 0.28
LM RMSE: 0.85
```

```
First 5 LM predictions(M-class):
0  0.008754
1 -0.016948
2  0.018749
3  0.053582
4  0.017597
```

```
M-class
LM MAE: 0.04
LM RMSE: 0.27
```

```
First 5 LM predictions(X-class):
0 -0.001148
1  0.005444
2 -0.000424
3  0.005046
4 -0.000497
```

```
X-class
LM MAE: 0.01
LM RMSE: 0.03
```

4. Prediction Model

Gradient Boosting Regressor:

```
First 5 GBM predictions(C-class):
0  0.037785
1  0.078790
2  0.100362
3  0.364807
4  0.085915
```

```
C-class
GBM MAE: 0.28
GBM RMSE: 0.85
```

```
First 5 GBM predictions(M-class):
0 -0.008721
1 -0.003070
2  0.013753
3  0.067197
4  0.003520
```

```
M-class
GBM MAE: 0.04
GBM RMSE: 0.27
```

```
First 5 GBM predictions(X-class):
0 -0.000246
1 -0.001688
2 -0.000143
3 -0.000915
4 -0.000518
```

```
X-class
GBM MAE: 0.01
GBM RMSE: 0.03
```

4. Prediction Model

Logistic Regression:

```
First 5 LOG predictions(C-class):
0  0
1  0
2  0
3  0
4  0
```

```
C-class
GBM MAE: 0.28
GBM RMSE: 0.86
```

```
First 5 LOG predictions(C-class):
0  0
1  0
2  0
3  0
4  0
```

```
M-class
GBM MAE: 0.03
GBM RMSE: 0.27
```

```
First 5 LOG predictions(C-class):
0  0
1  0
2  0
3  0
4  0
```

```
X-class
GBM MAE: 0.00
GBM RMSE: 0.00
```

We may think the prediction data consists only truncated of values of 0.
But if we display the full prediction, there will be some 1 that appears, which makes sense since in the columns it is based there's mostly 0.

Thank You !