# Filter Methods

May 23, 2021

[2]:
```
import pandas as pd

dataset = pd.read_csv('dataset1.csv')
X= dataset.drop(columns='Result')
Y= dataset['Result']
dataset.head()
```

[2]:

| | Links_in_tags | Abnormal_URL | Submitting_to_email | SFH | Iframe | popUpWidnow | \ |
|---|---|---|---|---|---|---|---|
| 0 | 0 | -1 | 1 | 1 | -1 | 1 | |
| 1 | 1 | -1 | 1 | 1 | -1 | 1 | |
| 2 | 1 | -1 | 1 | 1 | -1 | 1 | |
| 3 | -1 | -1 | 1 | -1 | 1 | -1 | |
| 4 | 0 | -1 | 1 | -1 | 1 | -1 | |

| | on_mouseover | RightClick | Redirect | Result |
|---|---|---|---|---|
| 0 | 1 | 1 | -1 | 1 |
| 1 | 1 | 1 | -1 | 1 |
| 2 | 1 | 1 | 0 | 1 |
| 3 | -1 | -1 | -1 | -1 |
| 4 | -1 | -1 | -1 | -1 |

## 0.1 1 - Using Multicollinearity

Multicollinearity can be detected using a technique called the Variance Inflation Factor(VIF). Generally, a VIF above 10 indicates a high multicollinearity.

When the relationship among the exploratory variables is exact, then it is the problem of very high multicollinearity, which should be removed from the data when regression analysis is conducted.

[3]:
```
# Detecting Multicollinearity with VIF
from statsmodels.stats.outliers_influence import variance_inflation_factor

# VIF dataframe
vif_data = pd.DataFrame()
vif_data["feature"] = X.columns

# calculating VIF for each feature
vif_data["VIF"] = [variance_inflation_factor(X.values, i)
                          for i in range(len(X.columns))]
```

```
print(vif_data)
```

```
               feature       VIF
0        Links_in_tags  1.157014
1         Abnormal_URL  1.970926
2  Submitting_to_email  7.906805
3                  SFH  5.584140
4               Iframe  7.034234
5          popUpWidnow  4.378522
6         on_mouseover  8.181712
7           RightClick  6.212225
8             Redirect  2.302129
```

when the VIF is greater than 10, then there is a problem of multicollinearity. How ever the above data shows that there is no VIF value greater than 10.

[ ]:

[ ]:

### 0.2 2 - Using Mean absolute Difference (MAD)

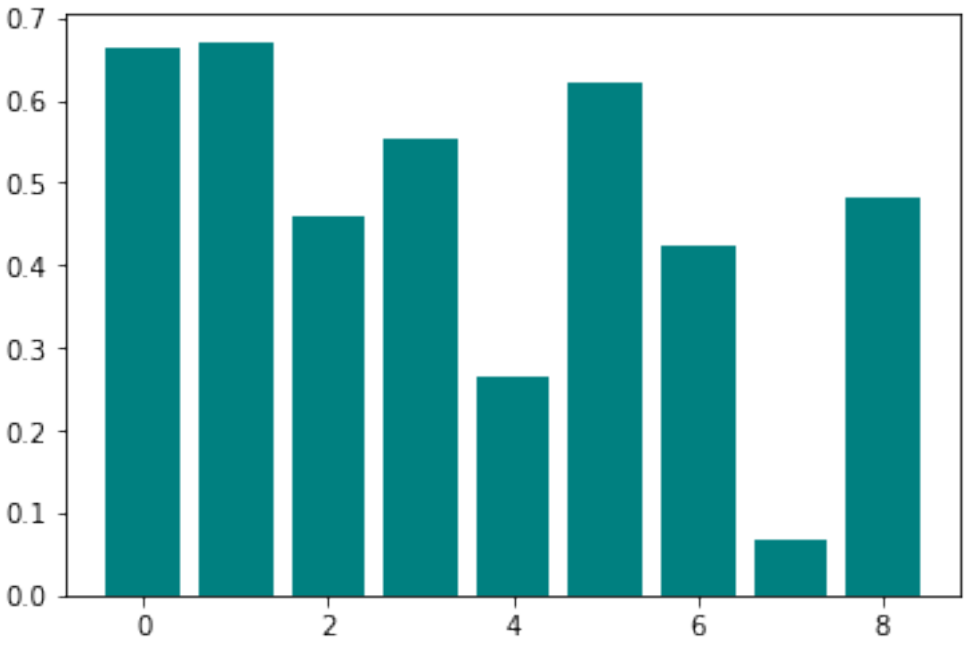
The mean absolute difference (MAD) computes the absolute difference from the mean value. The main difference between the variance and MAD measures is the absence of the square in the latter. The MAD, like the variance, is also a scale variant. This means that higher the MAD, higher the discriminatory power.

[5]:
```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

# Mean absolute Difference (MAD)
mean_abs_diff = np.sum(np.abs(X -np.mean(X, axis=0)), axis=0)/X.shape[0]

plt.bar(np.arange(X.shape[1]), mean_abs_diff, color='teal')
```

[5]: <BarContainer object of 9 artists>




According to above graph, we can see that lower MAD value is available for “RightCLick” feature. (that value is nearly 0.1)

[ ]: