Information Gain

May 23, 2021

Author: Lakruwan

The information gain of each attribute is calculated to find whether each attribute provides enough information for the prediction of the target/class. Some attributes may have same data in all over the dataset. Therefore those attributes may not be usefull.

Lets find out the information gain

[16]:
```
import pandas as pd

dataset = pd.read_csv('dataset1.csv')
dataset.drop('url', axis=1, inplace=True)
X= dataset.drop(columns='Result')
Y= dataset['Result']
X.head()
```

[16]:

	Links_in_tags	Abnormal_URL	Submitting_to_email	SFH	Iframe	popUpWidnow	on_mouseover	RightClick	Redirect
0	73.913043	-1	1	1	-1	1	1	1	0
1	85.000000	-1	1	1	-1	1	1	1	0
2	97.000000	-1	1	1	-1	1	1	1	1
3	12.000000	-1	1	-1	1	-1	-1	-1	0
4	55.555556	-1	1	-1	1	-1	-1	-1	0

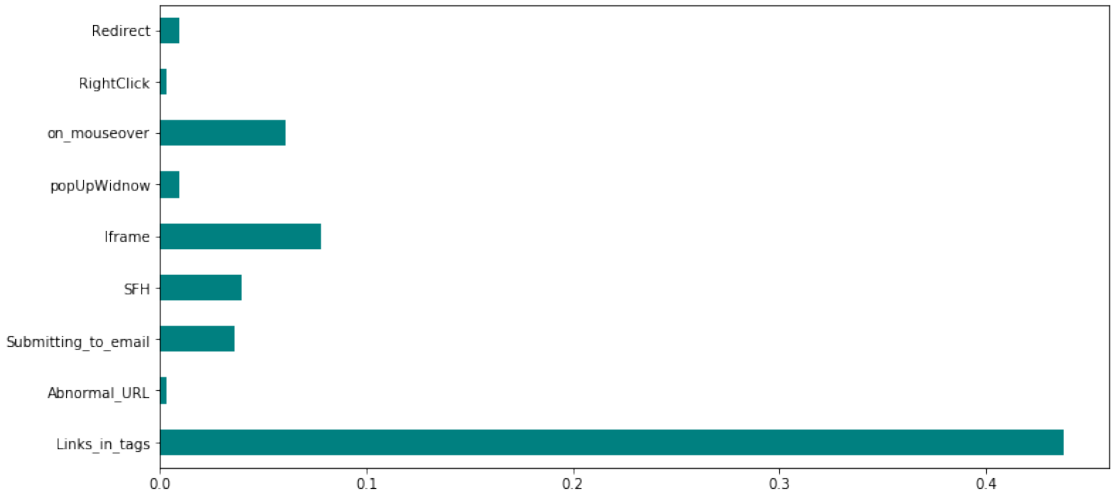
[29]:
```
# Information Gain
from sklearn.feature_selection import mutual_info_classif
import matplotlib.pyplot as plt
%matplotlib inline

importances = mutual_info_classif(X,Y) # here the mutual information between␣
 ↪variable and class is measured
info_gain = pd.Series(importances, dataset.columns[0:len(dataset.columns)-1])
info_gain
```

[29]:
```
Links_in_tags          0.437790
Abnormal_URL           0.002822
Submitting_to_email    0.036007
SFH                    0.039740
Iframe                 0.078289
popUpWidnow            0.009233
on_mouseover           0.060640
RightClick             0.003304
Redirect               0.009452
dtype: float64
```

[30]:
```
info_gain.plot(kind='barh', color='teal', figsize=(12,6))
```

[30]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1e95c3eb608>
```



According to above observation, links_in_tags feature provides higher amount of information for ditermination of class. And also other features also provide considerable amount of information for the class ditermination.

[ ]: