Preprocessing for CropMapping

May 5, 2021

[1]:
```
import pandas as pd
crop = pd.read_csv('Crop_mapping.txt', delimiter = ",")
crop.shape
```

[1]:
```
(325834, 175)
```

The dataset has a imbalanced distribution in the target variable ('label'):

[2]:
```
crop.label.value_counts()
```

[2]:
```
6    85074
3    75673
4    74067
5    47117
1    39162
2     3598
7     1143
Name: label, dtype: int64
```

Given that the dataset has a relatively high number of observations, the majority classes are downsampled to arrive at a balanced dataset.

[3]:
```
group = crop.groupby('label', group_keys=False)
balanced_crop = pd.DataFrame(group.apply(lambda x: x.sample(group.size().
↪min()))).reset_index(drop=True)
```

The dataset after downsampling now has a balanced label distribution:

[4]:
```
balanced_crop.label.value_counts()
```

[4]:
```
4    1143
1    1143
5    1143
2    1143
6    1143
3    1143
7    1143
Name: label, dtype: int64
```

[5]:
```
balanced_crop.shape
```

[5]:
```
(8001, 175)
```

[ ]:
```
# export the final dataframe to csv, which will be used in
# the assignment notebook

balanced_crop.to_csv('Crop_mapping.csv', index=False)
```