Exploratory Data Analysis - Detecting Class Imbalance ¶

In [1]:

```
import os
import matplotlib.pyplot as plt
```

In [2]:

```
!mkdir Train_
!mkdir Test_

!mkdir Train_/Fake
!mkdir Train_/Real

!mkdir Test_/Fake
!mkdir Test_/Real
```

In [3]:

```
# Repositioning the train data

PATH = "/kaggle/input/signature-verification/sign_data/train/"
for i in os.listdir(PATH):
    contol = i.split("_")
    try:
        if contol[1]=="forg":
            os.system("cp -r {} Train_/Fake".format(PATH+i))
    except:
        os.system("cp -r {} Train_/Real".format(PATH+i))
```

In [4]:

```
# Rearrange the test data

PATH = "/kaggle/input/signature-verification/sign_data/test/"

for i in os.listdir(PATH):
    contol = i.split("_")
    try:
        if contol[1]=="forg":
            os.system("cp -r {} Test_/Fake".format(PATH+i))
    except:
        os.system("cp -r {} Test_/Real".format(PATH+i))
```

In [5]:

```
# # loading training data
train_path = os.path.join("/kaggle/working/Train_")
# # loading testing data
test_path = os.path.join("/kaggle/working/Test_")
```

In [6]:

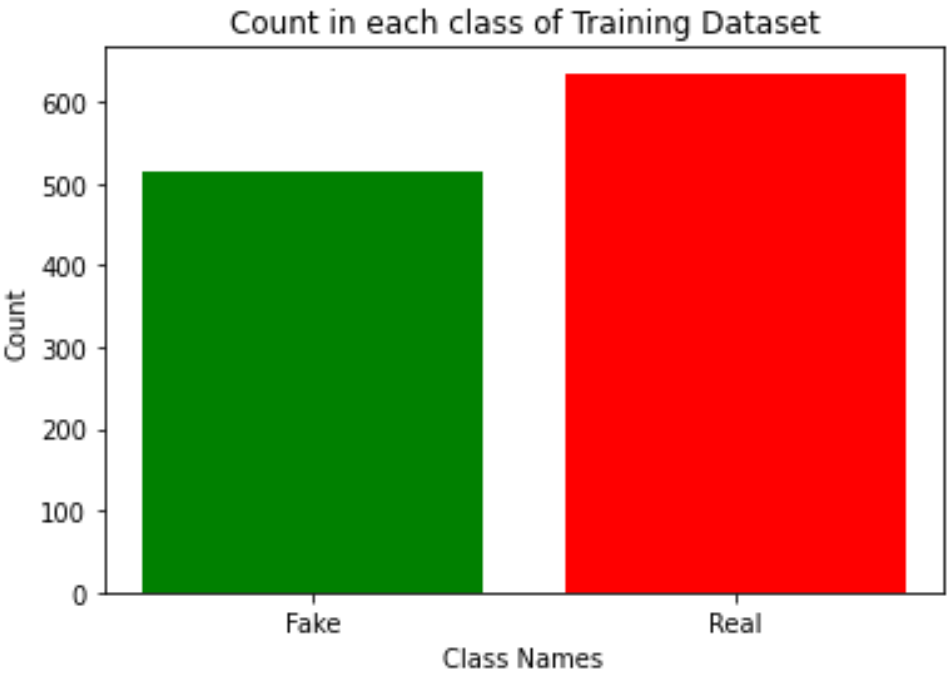
```
# Class imbalance in train data
train_fake = os.listdir(train_path+'/Fake')
train_real = os.listdir(train_path+'/Real')
fake_count = 0
real_count = 0
for folder in train_fake:
    fake_count += len(os.listdir(train_path +'/Fake/'+ folder))

for folder in train_real:
    real_count += len(os.listdir(train_path +'/Real/'+ folder))
total_count = real_count + fake_count
print('Fake img percentage: ', round(fake_count * 100/total_count))
print('Real img percentage: ', round(real_count * 100/total_count))
plt.bar( x = ['Fake', 'Real'], height = [fake_count, real_count], colo
r = ['green','red'])
plt.xlabel('Class Names')
plt.ylabel('Count')
plt.title('Count in each class of Training Dataset')
```

```
Fake img percentage:  45
Real img percentage:  55
```

Out[6]:

```
Text(0.5, 1.0, 'Count in each class of Training Dataset')
```



In [7]:

```
# Class Imbalance in test data
# This is done to use divide test data into equal sets of fake and real
 signatures
test_fake = os.listdir(test_path+'/Fake')
test_real = os.listdir(test_path+'/Real')
fake_count = 0
real_count = 0

for folder in test_fake:
    fake_count += len(os.listdir(test_path +'/Fake/'+ folder))

for folder in test_real:
    real_count += len(os.listdir(test_path +'/Real/'+ folder))
total_count = real_count + fake_count
print('Fake img percentage: ', round(fake_count * 100/total_count))
print('Real img percentage: ', round(real_count * 100/total_count))

plt.bar( x = ['Fake', 'Real'], height = [fake_count, real_count], colo
r = ['green','red'])
plt.xlabel('Class Names')
plt.ylabel('Count')
plt.title('Count in each class of Testing Dataset')
```

```
Fake img percentage:  50
Real img percentage:  50
```

Out[7]:

```
Text(0.5, 1.0, 'Count in each class of Testing Dataset')
```