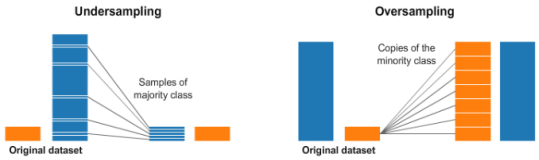
Step 1: RESAMPLING



Under- and Over-Sampling

1. Imports

In [1]:
```
import pandas as pd
```

2. Description

In our dataset we have the following data :

In [2]:
```
BStratosphere = pd.read_csv(r"C:\Users\Usuario\Documents\Github\PDG\PDG-2\Datasets\Time Window\Originals\Be
```

In [3]:
```
BStratosphere.shape
```

```
(519, 31)
```

In [4]:
```
MStratosphere = pd.read_csv(r"C:\Users\Usuario\Documents\Github\PDG\PDG-2\Datasets\Time Window\Originals\Ma
                            delimiter = ",")
```

In [5]:
```
MStratosphere.shape
```

```
(258178, 31)
```

As we can see, we have only **519** benign data compared with the malign data that is very huge. The above shows that we have **IMBALANCED DATA**.

In a first step, we generate a sample of benign data, trying to make a **OVERSAMPLING** of this type of data. We got the following number of time windows:

In [6]:
```
BOurResearch = pd.read_csv(r"C:\Users\Usuario\Documents\Github\PDG\PDG-2\Datasets\Time Window\Originals\Ben
```

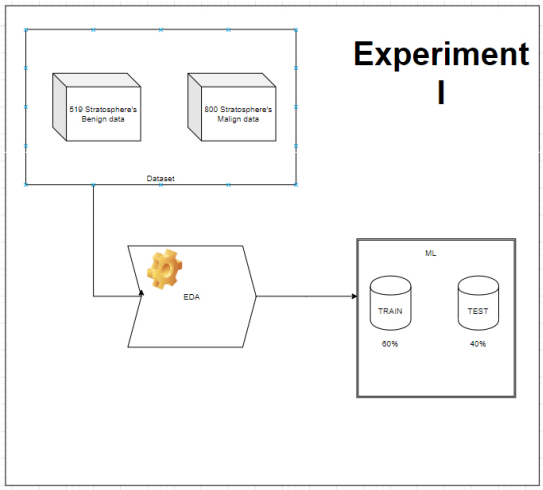
In [7]:
```
BOurResearch.shape
```

```
(649, 31)
```

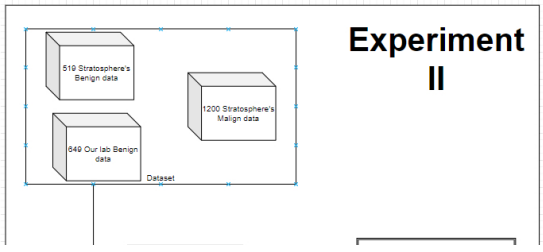
If we plus our data and the stratospheres benign data, we can get an approximate of **1.168** benign time windows. But, remains lower that malign data...

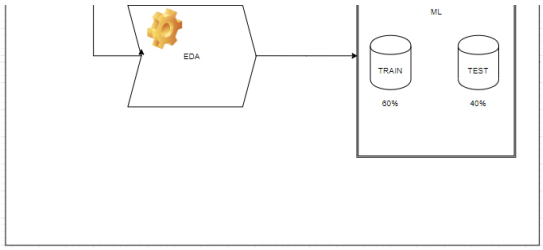
For this reason, we choose to apply a **UNDERSAMPLING** of **MALIGN DATA**, divided into **2 experiments**

EXPERIMENT -I



EXPERIMENT -II





REFERENCES

1. https://towardsdatascience.com/what-to-do-when-your-classification-dataset-is-imbalanced-6af031b12a36