CREATING THE FINAL DATASET

IMPORTING THE RAW DATASET

In [1]:

```
import pandas as pd
```

In [2]:

```
dataset = pd.read_csv('../dataset/project_dataset.csv', index_col = 0)
```

In [3]:

```
dataset.head()
```

Out[3]:

	Age	Gender	Polyuria	Polydipsia	Sudden weight loss	Weakness	Polyphagia	Genital thrush	Visual blurring	Itching	Irritability	Delayed healing	Partial paresis	Muscle stiffness	Alope
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	

ANALYSING THE DATASET

In [4]:

```
dataset.shape
```

Out[4]:

```
(640, 17)
```

In [5]:

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 640 entries, 0 to 639
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
```

```
---  ------              --------------  -----
 0   Age                 640 non-null    int64 
 1   Gender              640 non-null    object
 2   Polyuria            640 non-null    object
 3   Polydipsia          640 non-null    object
 4   Sudden weight loss  640 non-null    object
 5   Weakness            640 non-null    object
 6   Polyphagia          640 non-null    object
 7   Genital thrush      640 non-null    object
 8   Visual blurring     640 non-null    object
 9   Itching             640 non-null    object
 10  Irritability        640 non-null    object
 11  Delayed healing     640 non-null    object
 12  Partial paresis     640 non-null    object
 13  Muscle stiffness    640 non-null    object
 14  Alopecia            640 non-null    object
 15  Obesity             640 non-null    object
 16  Class               640 non-null    object
dtypes: int64(1), object(16)
memory usage: 90.0+ KB
```

In [6]:

```
dataset.nunique()
```

Out[6]:

```
Age                   50
Gender                 2
Polyuria               2
Polydipsia             2
Sudden weight loss     2
Weakness               2
Polyphagia             2
Genital thrush         2
Visual blurring        2
Itching                2
Irritability           2
Delayed healing        2
Partial paresis        2
Muscle stiffness       2
Alopecia               2
Obesity                2
Class                  2
dtype: int64
```

In [7]:

```
class_1 = dataset[dataset['Class'] == 'Positive']
class_0 = dataset[dataset['Class'] == 'Negative']
```

In [8]:

```
print("Number of positive outcomes :", len(class_1))
print("Number of negative outcomes :", len(class_0))
```

```
Number of positive outcomes : 320
Number of negative outcomes : 320
```

In [9]:

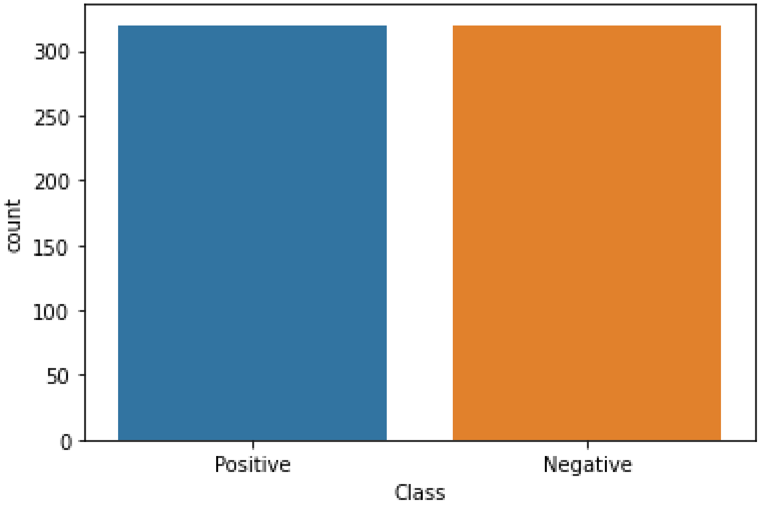
```
import seaborn as sns
```

In [10]:

```
sns.countplot(x = 'Class', data = dataset)
```

Out[10]:

```
<AxesSubplot:xlabel='Class', ylabel='count'>
```



USING LABELENCODER

In [11]:

```
from sklearn.preprocessing import LabelEncoder
labelencoder = LabelEncoder()
```

In [12]:

```
dataset['Gender'] = labelencoder.fit_transform(dataset['Gender'])
dataset['Polyuria'] = labelencoder.fit_transform(dataset['Polyuria'])
dataset['Polydipsia'] = labelencoder.fit_transform(dataset['Polydipsia'])
dataset['Sudden weight loss'] = labelencoder.fit_transform(dataset['Sudden weight loss'])
dataset['Weakness'] = labelencoder.fit_transform(dataset['Weakness'])
dataset['Polyphagia'] = labelencoder.fit_transform(dataset['Polyphagia'])
dataset['Genital thrush'] = labelencoder.fit_transform(dataset['Genital thrush'])
dataset['Visual blurring'] = labelencoder.fit_transform(dataset['Visual blurring'])
```

```
dataset['Itching'] = labelencoder.fit_transform(dataset['Itching'])
dataset['Irritability'] = labelencoder.fit_transform(dataset['Irritability'])
dataset['Delayed healing'] = labelencoder.fit_transform(dataset['Delayed healing'])
dataset['Partial paresis'] = labelencoder.fit_transform(dataset['Partial paresis'])
dataset['Muscle stiffness'] = labelencoder.fit_transform(dataset['Muscle stiffness'])
dataset['Alopecia'] = labelencoder.fit_transform(dataset['Alopecia'])
dataset['Obesity'] = labelencoder.fit_transform(dataset['Obesity'])
dataset['Class'] = labelencoder.fit_transform(dataset['Class'])
```

ANALYSING THE UPDATED DATASET

In [13]:

```
dataset.head()
```

Out[13]:

	Age	Gender	Polyuria	Polydipsia	Sudden weight loss	Weakness	Polyphagia	Genital thrush	Visual blurring	Itching	Irritability	Delayed healing	Partial paresis	Muscle stiffness	Alope
0	40	1	0	1	0	1	0	0	0	1	0	1	0	1	
1	58	1	0	0	0	1	0	0	1	0	0	0	1	0	
2	41	1	1	0	0	1	1	0	0	1	0	1	0	1	
3	45	1	0	0	1	1	1	1	0	1	0	1	0	0	
4	60	1	1	1	1	1	1	0	1	1	1	1	1	1	

In [14]:

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 640 entries, 0 to 639
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 640 non-null    int64
 1   Gender              640 non-null    int32
 2   Polyuria            640 non-null    int32
 3   Polydipsia          640 non-null    int32
 4   Sudden weight loss  640 non-null    int32
 5   Weakness            640 non-null    int32
 6   Polyphagia          640 non-null    int32
 7   Genital thrush      640 non-null    int32
 8   Visual blurring     640 non-null    int32
 9   Itching             640 non-null    int32
 10  Irritability        640 non-null    int32
 11  Delayed healing     640 non-null    int32
 12  Partial paresis     640 non-null    int32
 13  Muscle stiffness    640 non-null    int32
 14  Alopecia            640 non-null    int32
 15  Obesity             640 non-null    int32
 16  Class               640 non-null    int32
dtypes: int32(16), int64(1)
memory usage: 66.2 KB
```

SAVING THE FINAL DATASET

In [15]:

```python
import os.path
```

In [16]:

```python
if os.path.isfile('../dataset/train_dataset.csv') is False:
    dataset.to_csv('../dataset/train_dataset.csv')
```

In [ ]: