

BREAST CANCER DATASET EXPLORATORY DATA ANALYSIS PROJECT

Objective :

1) To get statistical insight of the dataset. 2) Checking the distribution of target variable. 3) Encoding the target column. 4) Grouping the data based on the target.

Tools to be used

1) NumPy 2) Pandas 3) LabelEncoder

Importing the Dependencies

```
In [1]: import numpy as np
import pandas as pd
from sklearn.preprocessing import LabelEncoder
```

The Dataset

```
In [2]: breast_cancer_data = pd.read_csv(r'C:\Users\HP\Downloads\data (1).csv')
```

```
In [3]: breast_cancer_data.head()
```

```
Out[3]:
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	pt
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	

5 rows × 33 columns

```
In [4]: breast_cancer_data.shape
```

```
Out[4]: (569, 33)
```

This dataset has 569 rows and 33 columns.

Checking if there is any null value in the dataset

```
In [5]: breast_cancer_data.isnull().sum()
```

```
Out[5]: id 0
diagnosis 0
radius_mean 0
texture_mean 0
perimeter_mean 0
area_mean 0
smoothness_mean 0
compactness_mean 0
concavity_mean 0
concave points_mean 0
symmetry_mean 0
fractal_dimension_mean 0
radius_se 0
texture_se 0
perimeter_se 0
area_se 0
smoothness_se 0
compactness_se 0
concavity_se 0
concave points_se 0
symmetry_se 0
fractal_dimension_se 0
radius_worst 0
texture_worst 0
perimeter_worst 0
area_worst 0
smoothness_worst 0
compactness_worst 0
concavity_worst 0
concave points_worst 0
symmetry_worst 0
fractal_dimension_worst 0
Unnamed: 32 569
dtype: int64
```

This dataset has no null values.

Checking the data types

```
In [6]: breast_cancer_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 33 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   id                                     569 non-null    int64
1   diagnosis                             569 non-null    object
2   radius_mean                           569 non-null    float64
3   texture_mean                           569 non-null    float64
4   perimeter_mean                         569 non-null    float64
5   area_mean                             569 non-null    float64
6   smoothness_mean                        569 non-null    float64
7   compactness_mean                       569 non-null    float64
8   concavity_mean                         569 non-null    float64
9   concave points_mean                   569 non-null    float64
10  symmetry_mean                          569 non-null    float64
11  fractal_dimension_mean                 569 non-null    float64
12  radius_se                              569 non-null    float64
13  texture_se                             569 non-null    float64
14  perimeter_se                           569 non-null    float64
15  area_se                                569 non-null    float64
16  smoothness_se                          569 non-null    float64
17  compactness_se                         569 non-null    float64
18  concavity_se                           569 non-null    float64
19  concave points_se                      569 non-null    float64
20  symmetry_se                            569 non-null    float64
21  fractal_dimension_se                   569 non-null    float64
22  radius_worst                           569 non-null    float64
23  texture_worst                           569 non-null    float64
24  perimeter_worst                        569 non-null    float64
25  area_worst                             569 non-null    float64
26  smoothness_worst                       569 non-null    float64
27  compactness_worst                      569 non-null    float64
28  concavity_worst                        569 non-null    float64
29  concave points_worst                   569 non-null    float64
30  symmetry_worst                         569 non-null    float64
31  fractal_dimension_worst                 569 non-null    float64
32  Unnamed: 32                             0 non-null      float64
dtypes: float64(31), int64(1), object(1)
memory usage: 146.8+ KB
```

Removing the 'Unnamed: 32' column from the dataframe

```
breast_cancer_data.drop(columns='Unnamed: 32', axis=1, inplace=True)
```

```
In [7]: breast_cancer_data.drop(columns='Unnamed: 32',axis=1,inplace=True)
```

```
In [8]: breast_cancer_data.shape
```

```
Out[8]: (569, 32)
```

```
In [9]: breast_cancer_data.head()
```

```
Out[9]:
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	points_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	147.10
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	107.01
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	127.90
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	105.20
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	104.30

5 rows × 32 columns

Removing the 'id' column from the dataset

```
In [10]: breast_cancer_data.drop(columns='id',axis=1,inplace=True)
```

```
In [11]: breast_cancer_data.shape
```

```
Out[11]: (569, 31)
```

```
In [12]: breast_cancer_data.head()
```

```
Out[12]:
```

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
0	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710
1	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017
2	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790
3	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520
4	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430

5 rows × 31 columns

Checking for any null value

```
In [13]: breast_cancer_data.isnull().sum()
```

```
Out[13]: diagnosis                0
radius_mean                    0
texture_mean                   0
perimeter_mean                 0
area_mean                     0
smoothness_mean               0
compactness_mean              0
concavity_mean                0
concave points_mean           0
symmetry_mean                 0
fractal_dimension_mean        0
radius_se                     0
texture_se                    0
perimeter_se                  0
area_se                       0
smoothness_se                 0
compactness_se                0
concavity_se                  0
concave points_se             0
symmetry_se                   0
fractal_dimension_se          0
radius_worst                  0
texture_worst                 0
perimeter_worst               0
area_worst                    0
smoothness_worst              0
compactness_worst             0
concavity_worst               0
concave points_worst          0
symmetry_worst                0
fractal_dimension_worst       0
dtype: int64
```

This dataset has no null values.

Statistical insight of the dataset

```
In [14]: breast_cancer_data.describe()
```

```
Out[14]:
```

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	sym
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	

8 rows × 30 columns

Checking the distribution of target variable

```
In [15]: breast_cancer_data['diagnosis'].value_counts()
```

```
Out[15]: B      357
M      212
Name: diagnosis, dtype: int64
```

According to the given dataset there are 357 cases of benign breast cancer whereas 212 cases of malignant breast cancer.

Encoding the target column

```
In [16]: label_encode = LabelEncoder()
```

```
In [17]: labels=label_encode.fit_transform(breast_cancer_data['diagnosis'])
```

```
In [18]: breast_cancer_data['target'] = labels
```

```
In [19]: breast_cancer_data.drop(columns='diagnosis',axis=1,inplace=True)
```

```
breast_cancer_data.head()
```

```
In [21]: breast_cancer_data['target'].value_counts()
```

```
Out[21]: 0    357
         1    212
         Name: target, dtype: int64

         0 ----> Benign 1 ----> Malignant
```

Grouping the data based on the target

```
In [24]: breast_cancer_data.groupby('target').mean()
```

```
Out[24]:
```

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symn
target									
0	12.146524	17.914762	78.075406	462.790196	0.092478	0.080085	0.046058	0.025717	
1	17.462830	21.604906	115.365377	978.376415	0.102898	0.145188	0.160775	0.087990	

2 rows × 30 columns

It is clearly observed that the parameters are higher in case of malignant breast cancer than those of benign breast cancer.

Conclusion :

According to the given dataset there are 357 cases of benign breast cancer whereas 212 cases of malignant breast cancer. The value of various targets such as radius mean , texture mean , perimeter mean , area mean , smoothness mean etc. are higher in malignant cases than benign cases.

```
In [ ]:
```