Dev Joshi

Final Report : Benign or Malignant Cancer Prediction

Problem Statement

The dataset consisting of digitized imaging of fine needle aspirate (FNA) of a breast tumor cell mass can be used to distinguish between a benign and malignant tumor to aid in clinical diagnosis. Each cell nucleus has ten real-valued features which describe characteristics of the cell nuclei present in the image. The project goal is to deploy machine learning algorithms to accurately distinguish between a benign and malignant tumor to aid in clinical diagnosis. The original dataset is available at the https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+Diagnostic .

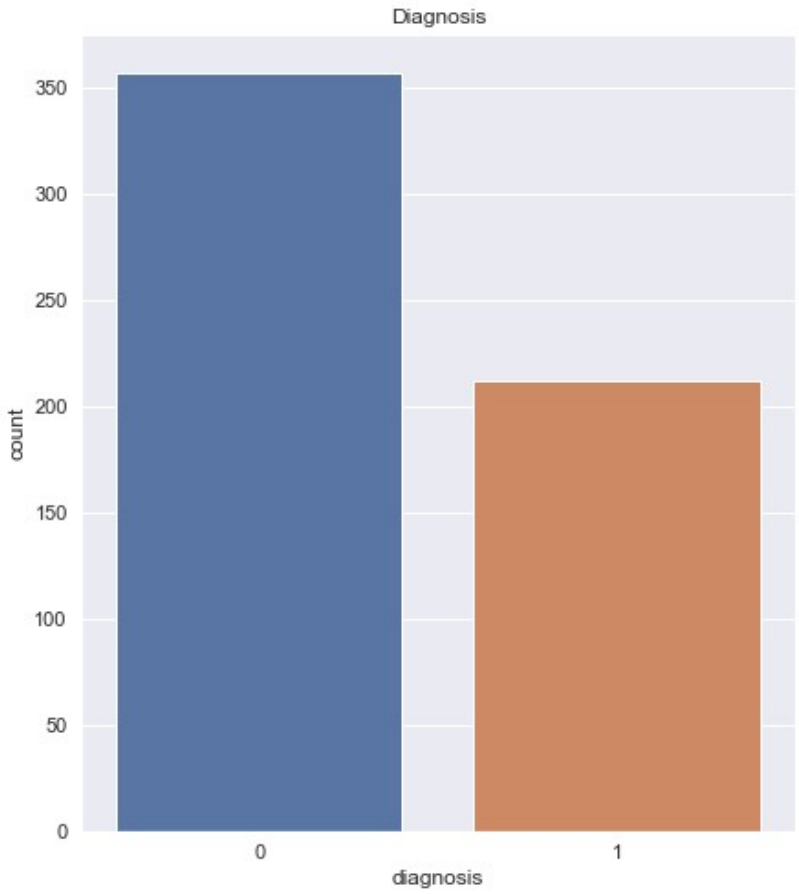
Data Wrangling: We began by downloading the dataset from the above link. The dataset consisted of 569 rows and 33 columns. We removed two columns – 'id' and 'Unnamed: 32' as they weren't essential for the analysis. The various columns of the dataset are :

```
0   diagnosis
1   radius_mean
2   texture_mean
3   perimeter_mean
4   area_mean
5   smoothness_mean
6   compactness_mean
7   concavity_mean
8   concave points_mean
9   symmetry_mean
10  fractal_dimension_mean
11  radius_se
12  texture_se
13  perimeter_se
14  area_se
15  smoothness_se
16  compactness_se
17  concavity_se
18  concave points_se
19  symmetry_se
20  fractal_dimension_se
21  radius_worst
22  texture_worst
23  perimeter_worst
```

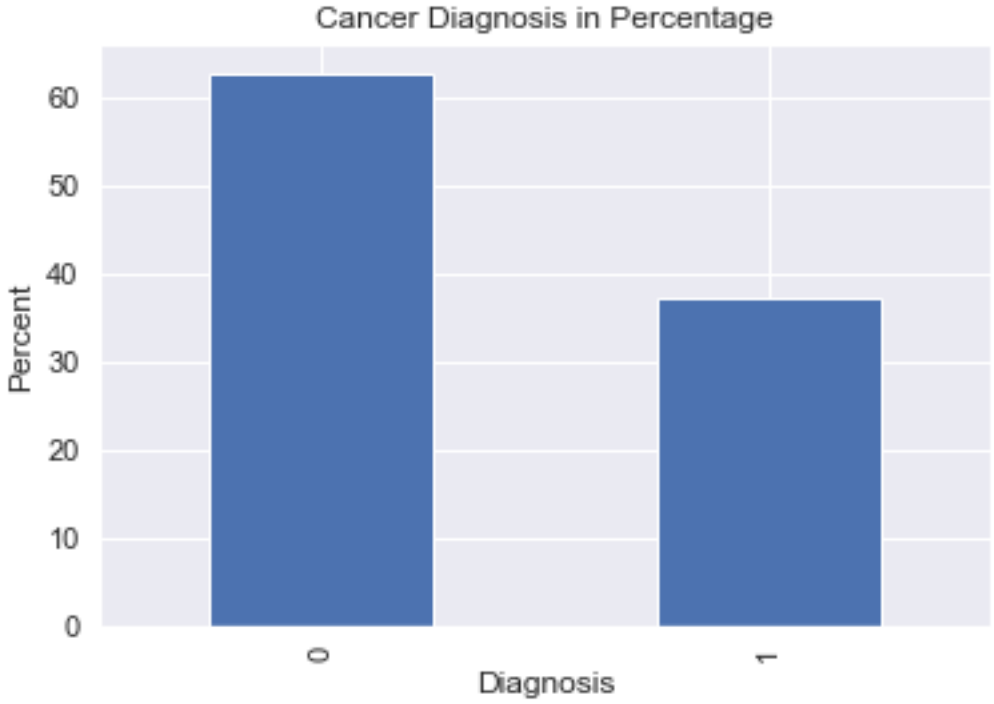
```
24  area_worst
25  smoothness_worst
26  compactness_worst
27  concavity_worst
28  concave points_worst
29  symmetry_worst
30  fractal_dimension_worst
```

We explored the data and looked at various columns of the dataframe. We made the target variable (column) 'diagnosis ' as numeric by assigning '1' as Malignant and '0' as benign.

We plotted this column to observe that the number of the benign cases are 357 and the number of malignant cases are 212.

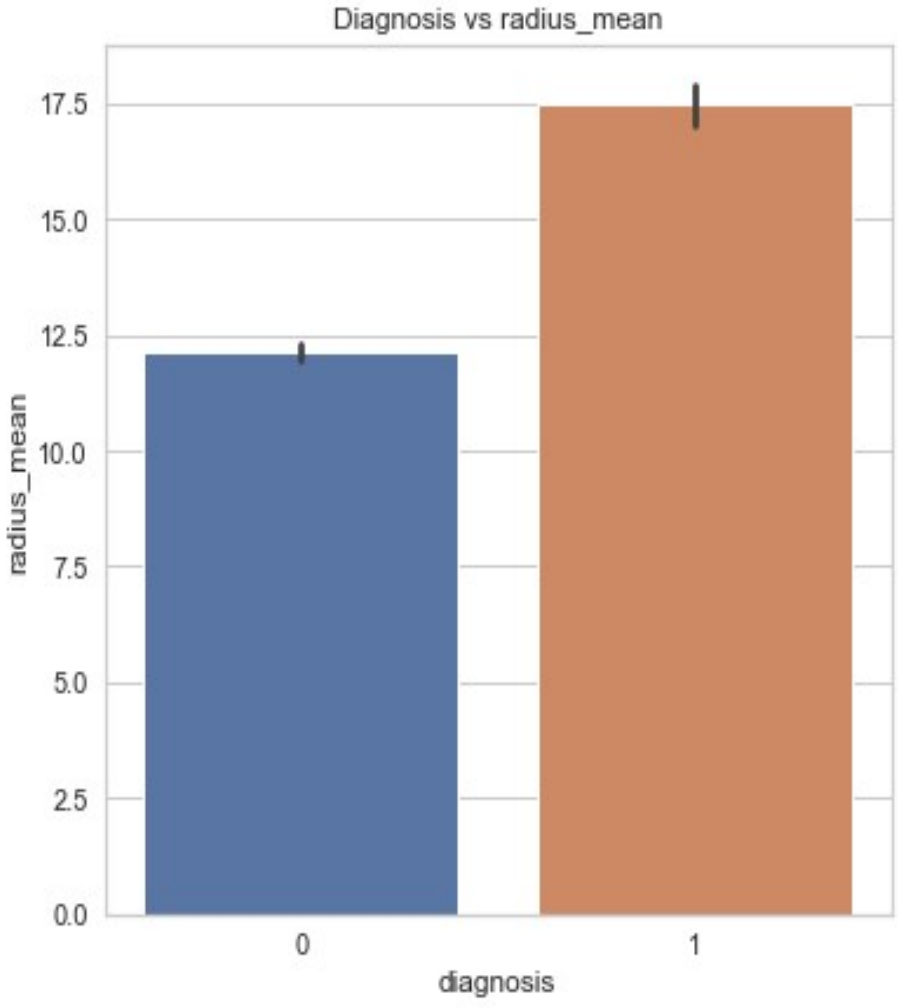


The same plot in terms of percentage yielded 62.7 % and 37.3 % for 0 (benign) and 1(malignant) respectively as seen in the follow image.



Exploratory Data Analysis:

After initial data wrangling, we sought to explore the various features of the data. We looked at the various tumor features in relation to the diagnosis. For example, we see that the malign cells have greater mean radius than benign cell as seen in the image aside.



Similarly, the other features in the data can also be compared as seen in the image below.