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 .

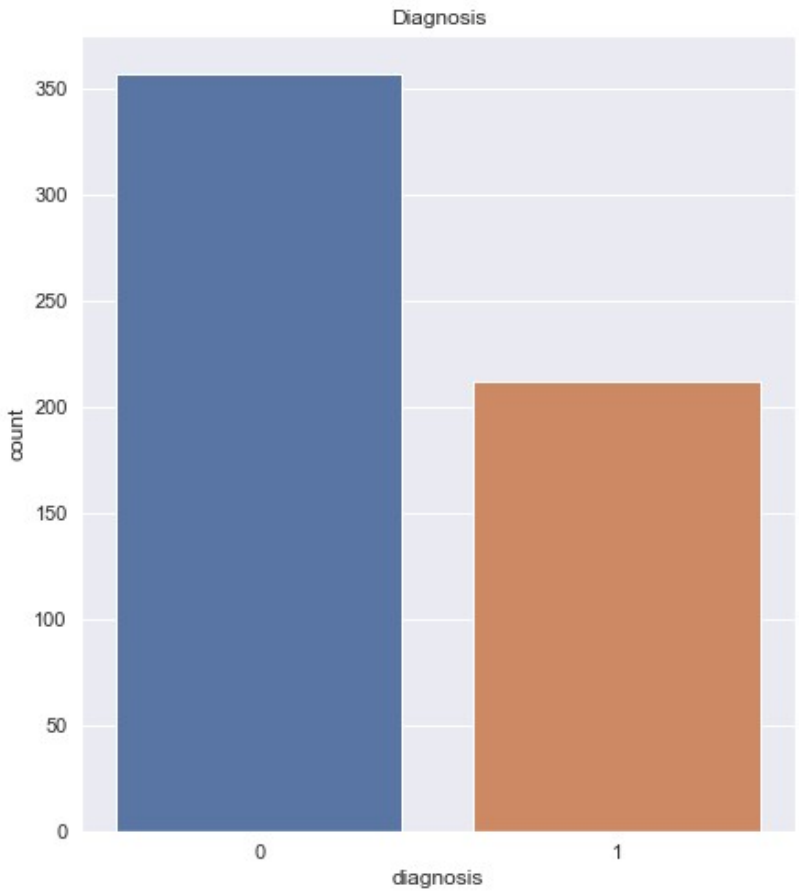
Data Wrangling: We began by downloading the dataset from the above link. The dataset consisted of 569 rows and 33 columns. We removed two columns – 'id' and 'Unnamed: 32' as they weren't essential for the analysis. The various columns of the dataset are :

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

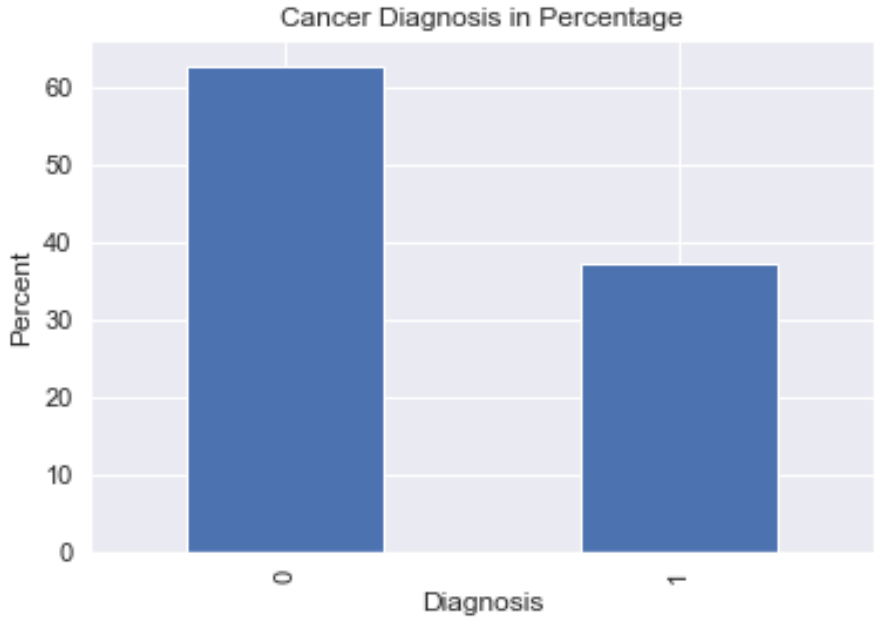
22 texture_worst
23 perimeter_worst
24 area_worst
25 smoothness_worst
26 compactness_worst
27 concavity_worst
28 concave points_worst
29 symmetry_worst
30 fractal_dimension_worst

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.

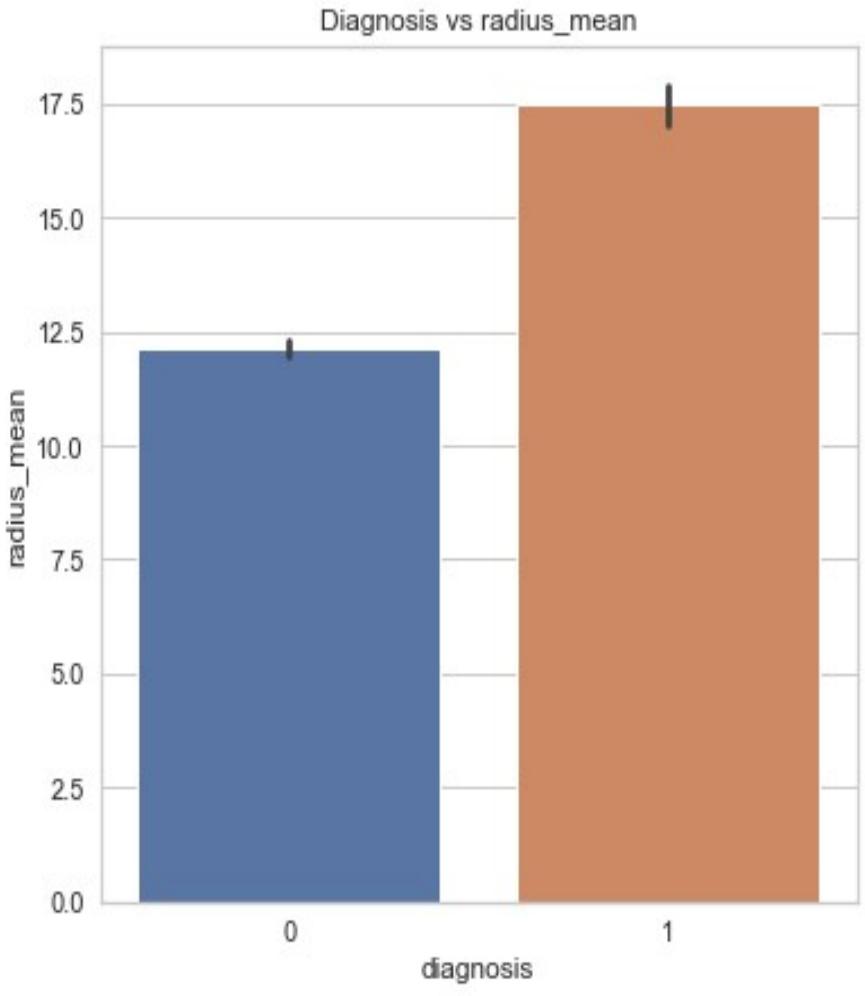
We plotted this column to observe that the number of the benign cases are 357 and the number of malignant cases are 212.



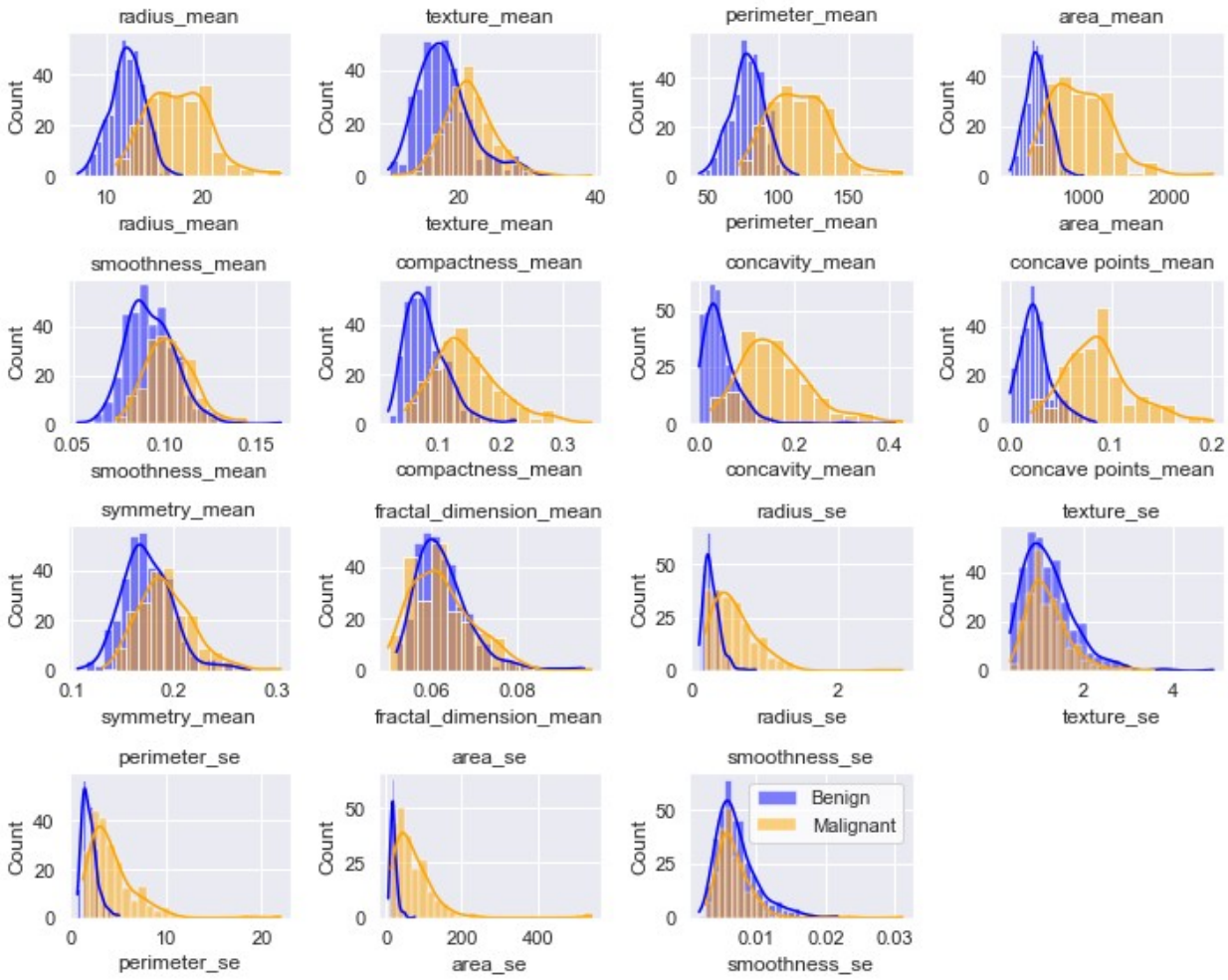
The same plot in terms of percentage yielded 62.7 % and 37.3 % for 0 (benign) and 1(malignant) respectively as seen in the following image.



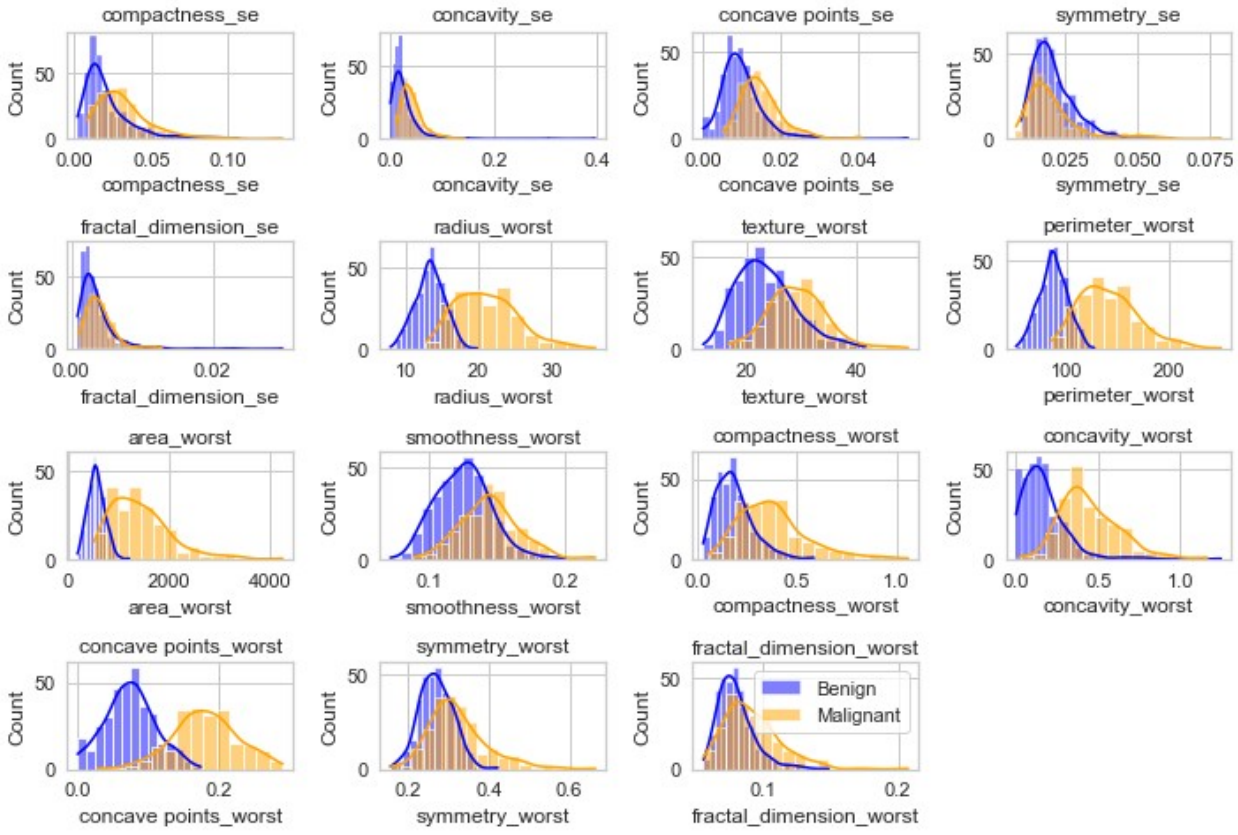
Exploratory Data Analysis: After initial data wrangling, we sought to explore the various features of the data. We looked at the various tumor features in relation to the diagnosis. For example, we see that the malign cells have greater mean radius than benign cell as seen in the adjacent image .



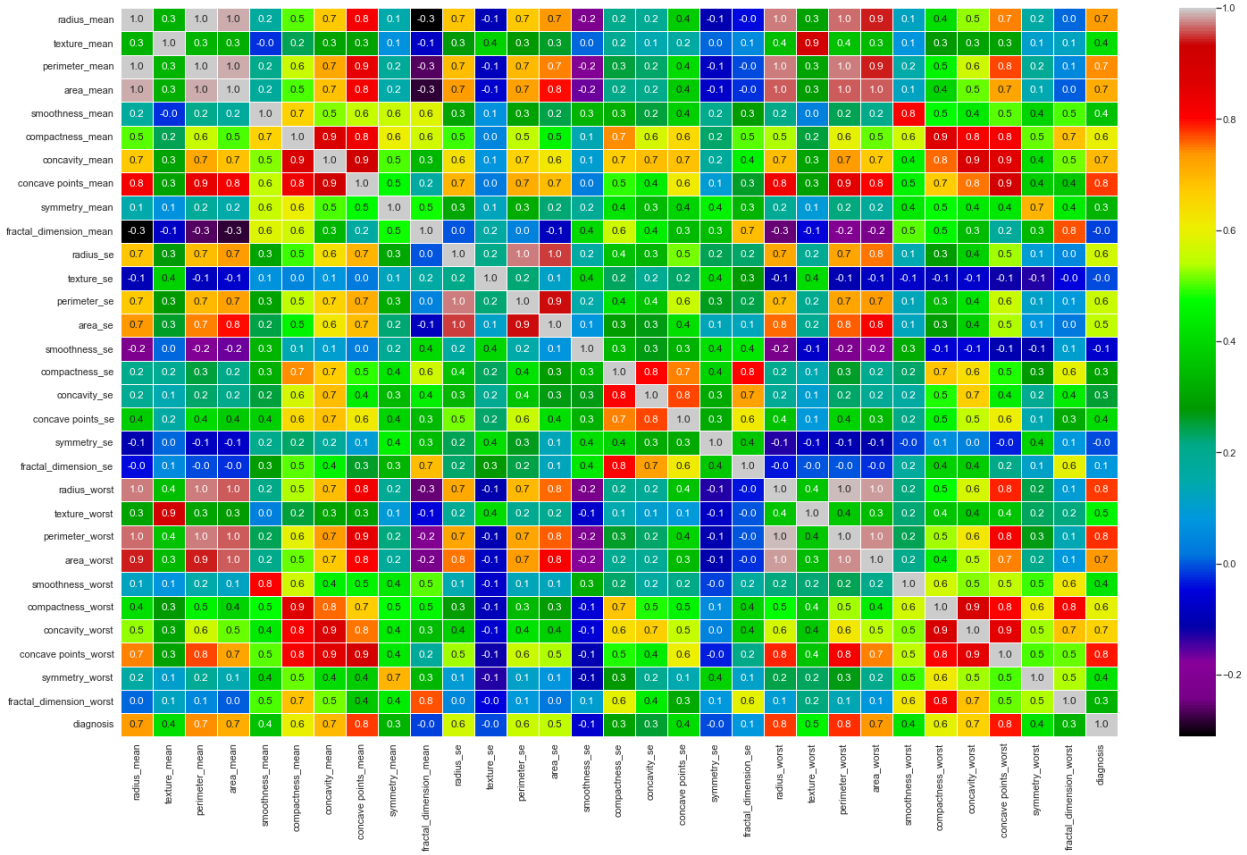
Similarly, the other features in the data can also be compared as seen in the image below. From this image also, we see that the mean radius is greater for the malign cells as compared to the benign cells (1st subplot). For most of the features (except few – namely – symmetry_mean, fractal_dimension_mean, texture_se, smoothness_se), the feature corresponding to the malign cells has higher value as compared to that of the benign cells.



We plotted similar image for the remaining features as well. Here also, we see the malign cells have higher value, although lesser count, than the benign cells for most of the features.

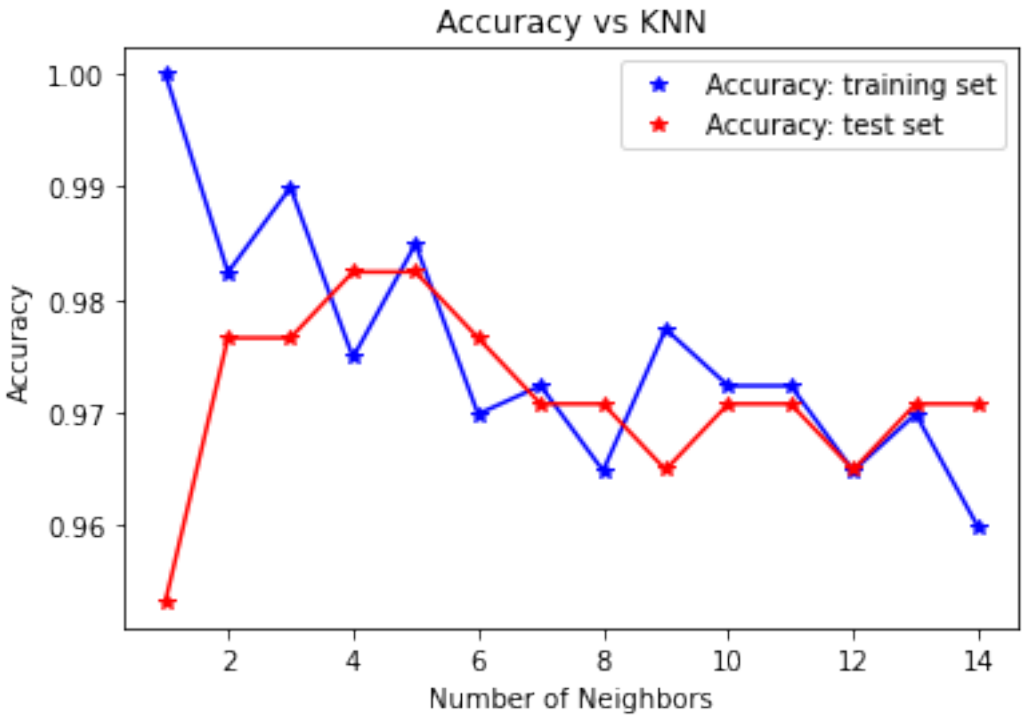


We also checked the correlation map with respect to the diagnosis feature. As we can see in the image below, it has a low correlation - but positive - with most of the features. It has negative correlation with few of the features. Lastly, we saved the data for further analysis in the subsequent units.



Preprocessing, Training and Modeling:

We downloaded the saved data from the previous section. We then divided the data into those that will be used to train the model and those that will be used to predict the approval : 70 % for training and 30 % for testing. We also applied standard scaling for X_train and X_test data so that no feature (with larger values) dominates over others (with smaller values).


Accuracy vs KNN
Accuracy: training set
Accuracy: test set
1.00
0.99
0.98
0.97
0.96
Accuracy
2
4
6
8
10
12
14
Number of Neighbors