
Breast Cancer Classification

Milestone Report

OVERVIEW

Breast cancer (BC) is one of the most common cancers among women in the world today. An early diagnosis of BC can greatly improve the prognosis and chance of survival for patients. An accurate classification of benign tumors can prevent patients from undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of tumors into malignant or benign groups is a subject of much ongoing research. With the use of advanced machine learning algorithms, I plan to build a model which accurately classifies tumors as Benign or Malignant based on certain features.

DATA

The dataset has been obtained from Kaggle. It contains 596 rows and 32 columns of tumor shape and specifications. The tumor is ultimately classified as benign or malignant based on its geometry and shape. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. There are 212 malignant and 357 benign tumors in the dataset.

The features of the dataset include:

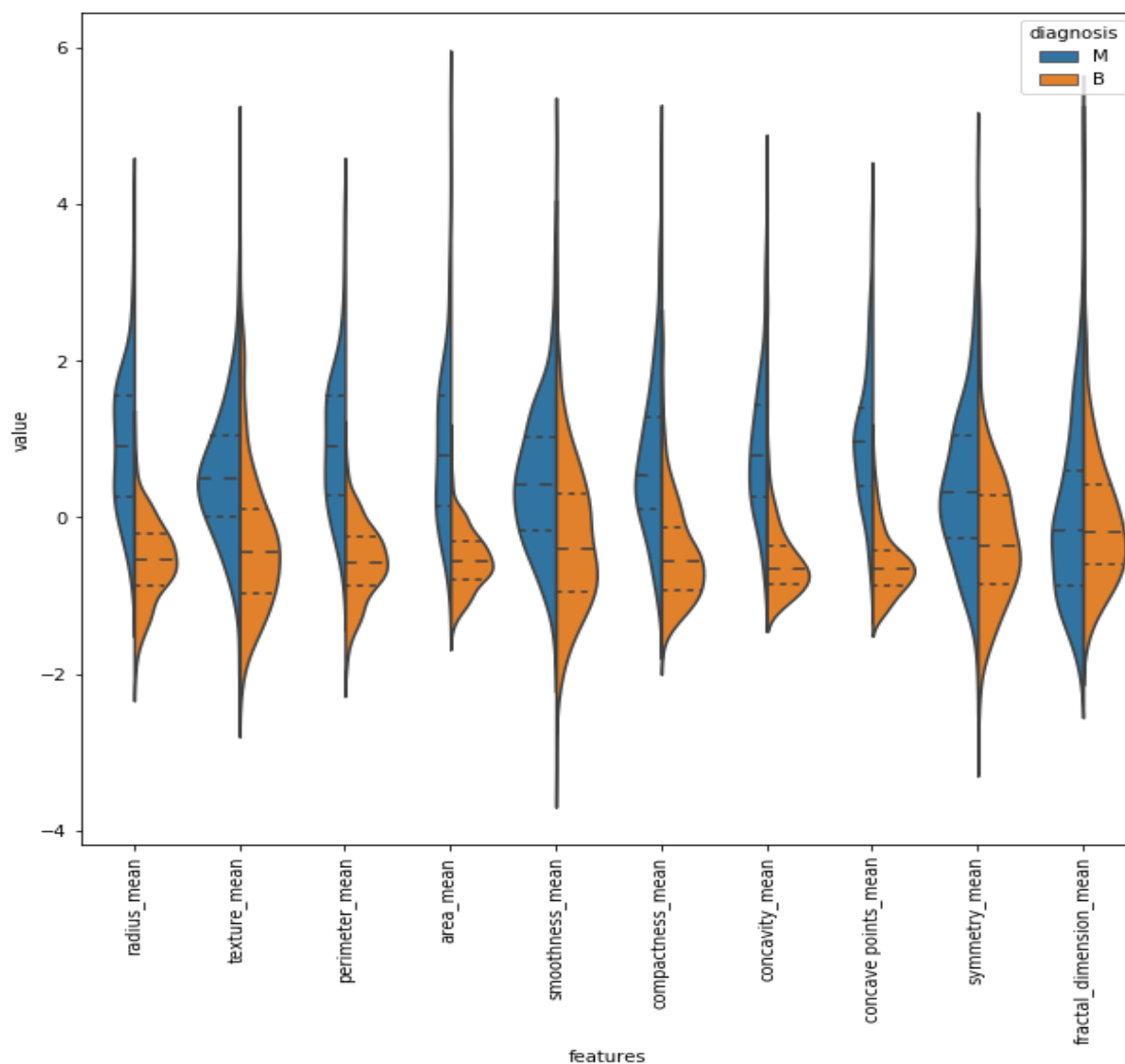
1. tumor radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness ($\text{perimeter}^2 / \text{area} - 1.0$)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.

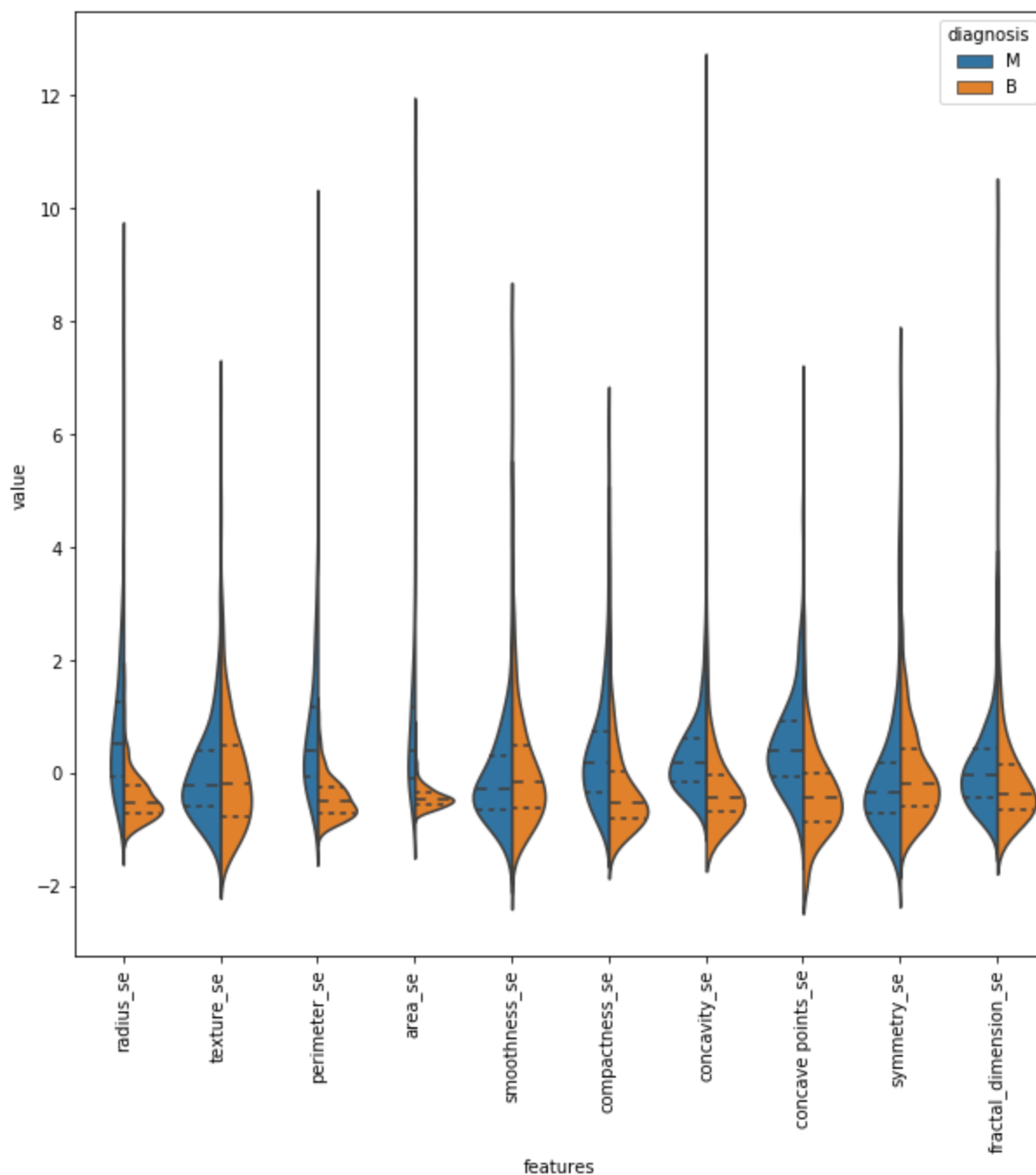
EXPLORATORY DATA ANALYSIS

Violin Plots

The dataset did not have any missing or null values. So I moved on to EDA and visualization. Violin plots were plotted to visualize the 30 features in order to get some insights into the distribution pattern of the features, their mean, std deviation or variance. A violin plot shows the distribution of quantitative data across several levels of one (or more) categorical variables such that those distributions can be compared. The mean, standard error and worst dimensions of the ten features were plotted separately in each series of violin plots.

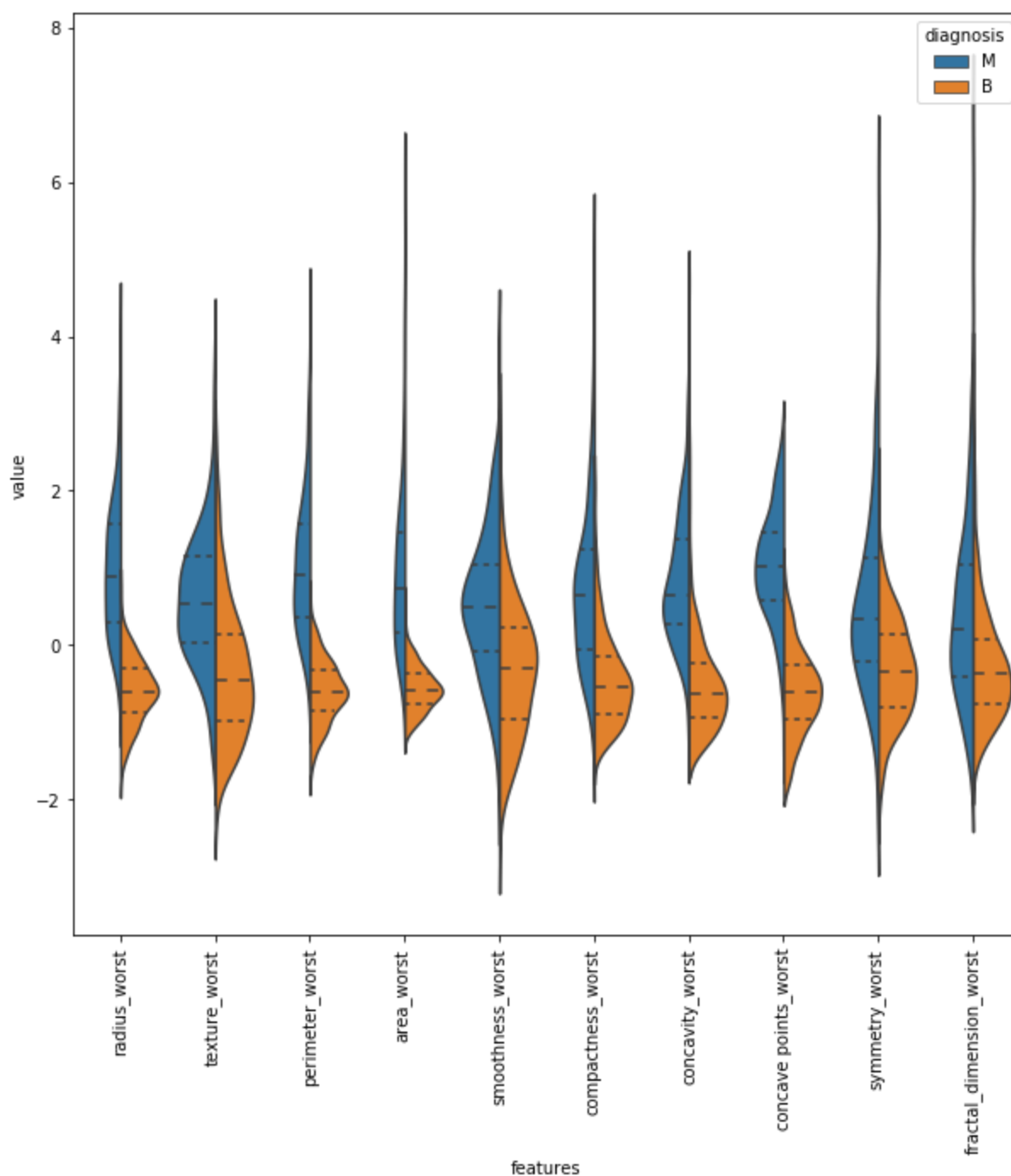


The median of texture_mean for Malignant and Benign looks separated, so it might be a good feature for classification. For fractal_dimension_mean, the medians of the Malignant and Benign groups are very close to each other which might not be good for classification. smoothness_mean seems to have the highest range of values.



The medians for almost all Malignant or Benign don't vary much for the standard error features above, except for concave points_se and concavity_se. smoothness_se or symmetry_se have a very similar distribution which could make classification using this

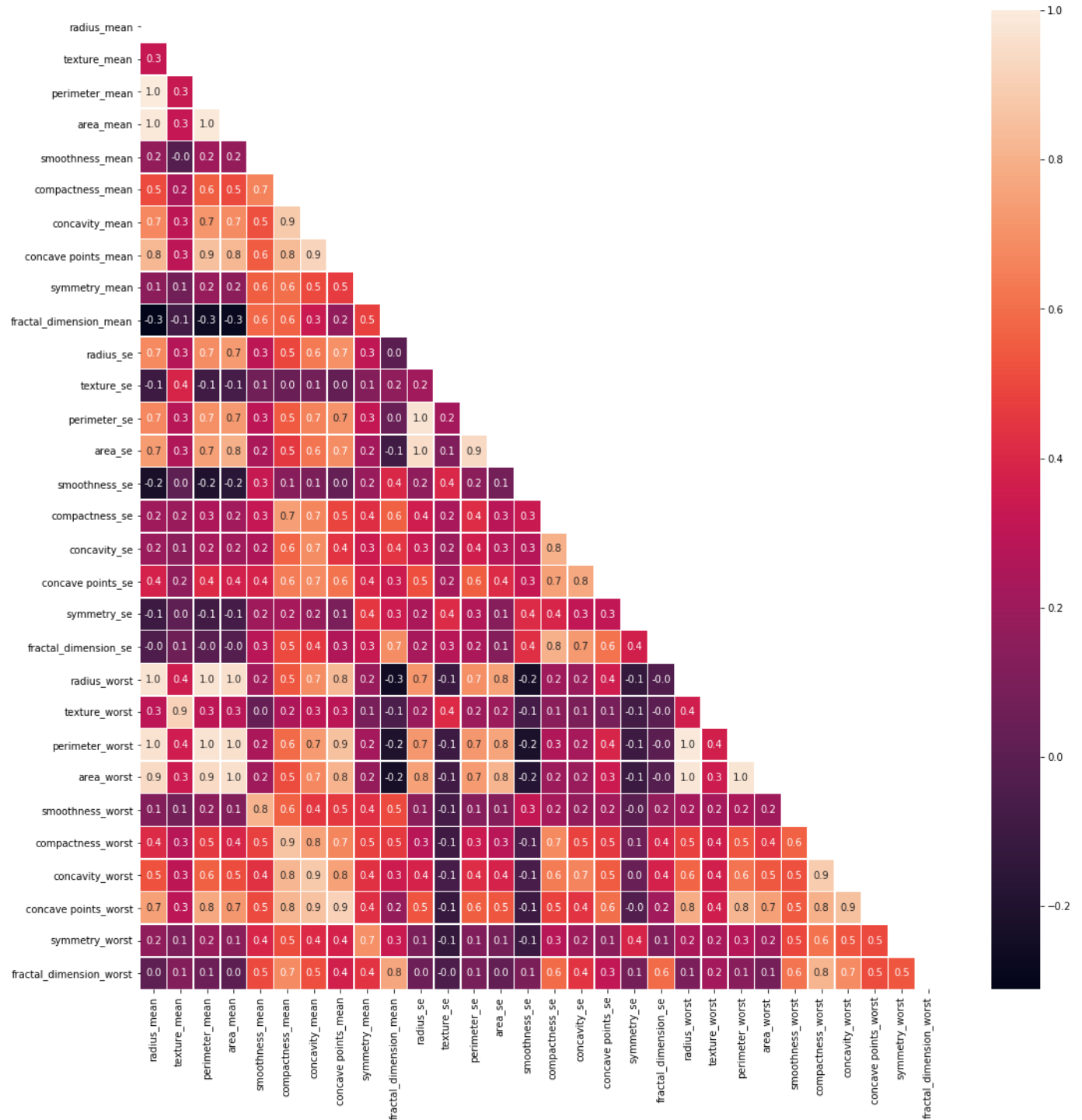
feature difficult. The shape of the violin plot for area_se looks warped. The distribution of data points for benign and malignant in area_se looks very different and varies the most.



area_worst look well separated, so it might be easier to use this feature for classification! Variance seems highest for fractal_dimension_worst. concavity_worst and concave_points_worst seem to have a similar data distribution. Let us check if these features are correlated.

Correlation Matrix

A correlation matrix was plotted to check for correlation and multicollinearity between the features.

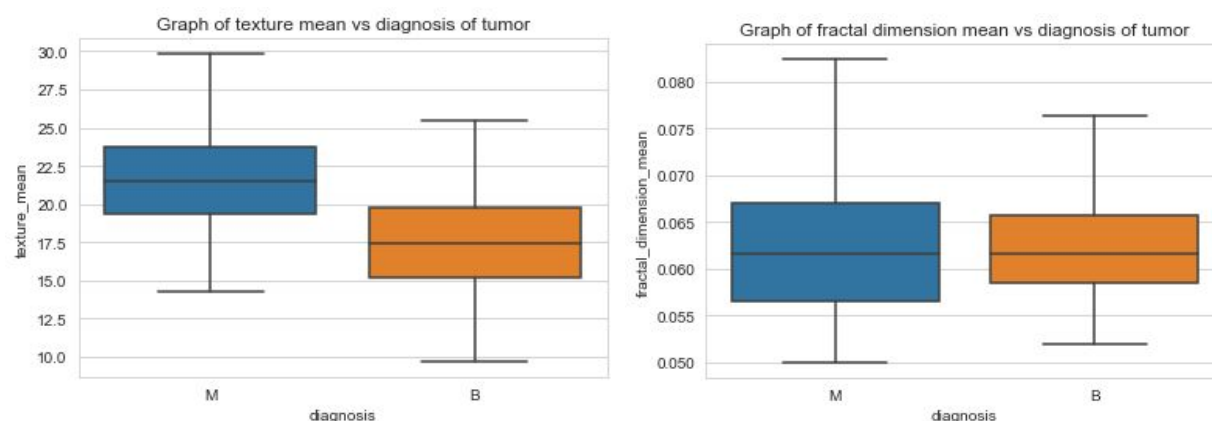


The means, std errors and worst dimension lengths of compactness, concavity and concave points of tumors are highly correlated amongst each other (correlation > 0.8). The std

errors of radius, perimeter and area of tumors have a correlation of 1! The worst dimension of radius, perimeter and area also have a correlation of 1. texture_mean and texture_worst have a correlation of 0.9. area_worst and area_mean have a correlation of 1. The mean and worst dimension of radius, perimeter and area have a correlation of 1.

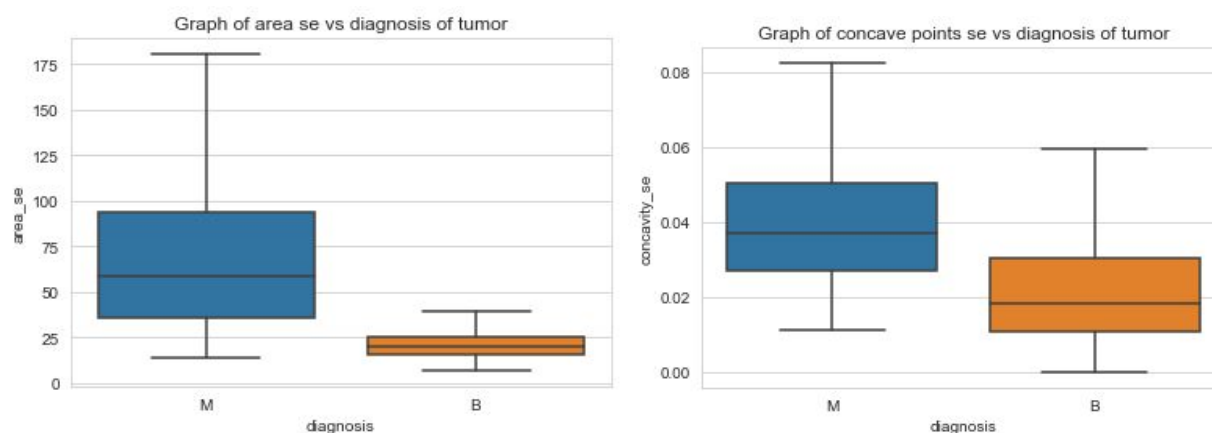
STATISTICAL ANALYSIS

Box Plots



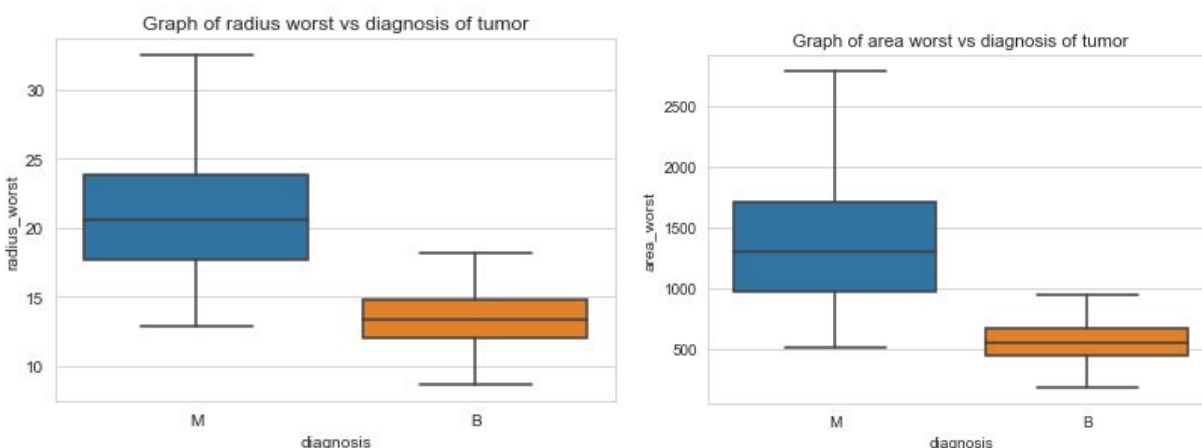
Texture means, for malignant and benign tumors vary by about 3 units. The distribution looks similar for both the groups. Malignant tumors tend to have a higher texture mean compared to benign.

Fractal dimension means are almost the same for malignant and benign tumors. The IQR is wider for malignant tumors.



Malignant groups have a distinctly wider range of values for area se. The distribution range is very narrow for benign groups. This might be a good feature for classification.

Standard error (se) of Concave points has a higher mean and IQR for malignant tumors. The distribution looks somewhat similar for both tumor types.




Malignant groups have a wider range of values for radius worst compared to benign groups. The IQR is wider for the same. Malignant tumors have a higher radius worst compared to benign groups.

Similar to area_se, area_worst has very different data distribution for malignant and benign tumors. Malignant tumors tend to have a higher value of mean and wider IQR range. The area_worst for benign tumors varies only between 10-1000, whereas the range is 500-2500 for malignant tumors. Because of noticeable differences between B and M tumors, this could be a good feature for classification.

t-test

| Features | t-statistic | p-value |
|------------------------|-------------|-----------------|
| texture mean | 10.86720108 | 4.05863605e-25 |
| fractal dimension mean | -0.30571113 | 0.7599368 |
| area se | 15.6093429 | 5.89552139e-46 |
| concave points se | 6.24615734 | 8.26017617e-10 |
| radius worst | 29.33908156 | 8.48229192e-116 |
| area worst | 25.72159026 | 2.8288477e-97 |

Except for fractal dimension mean, the p value and t statistic is statistically significant for all the features in the table above. The t statistic for the fractal dimension mean, is negative so if there is a difference between the M and B samples, it will be in the negative direction, meaning M tumor samples have lesser means than B tumor samples. However the value of t statistic is very small and p value > 0.01, this means we cannot reject null hypothesis. The difference in means for fractal dimension_mean samples of M and B tumors might not be statistically significant.



From the correlation matrix it was clear that there are quite a few features with very high correlations. So I dropped one of the features, from each of the feature pairs which had a correlation greater than 0.95. 'perimeter_mean', 'area_mean', 'perimeter_se', 'area_se', 'radius_worst', 'perimeter_worst', 'area_worst' were amongst the features that were dropped.

CONCLUSION

We are now ready to start machine learning on the dataset. I will try out some more feature selection techniques and ultimately build a classifier which can accurately predict whether a tumor is malignant or benign.