

BREAST CANCER DETECTION

Name: Sanjna

Roll No.: 52

Registration No. 11712337

Section: KM032

# ABOUT DATASET

This dataset is consist 12 features and each feature is containing different values like **radius, smoothness,** etc. They describe characteristics of the cell nuclei present in the image. It also shows certain parameter through which we can classify whether a person is having breast cancer or not. This dataset is having 569 instances. It is having 32 attributes.

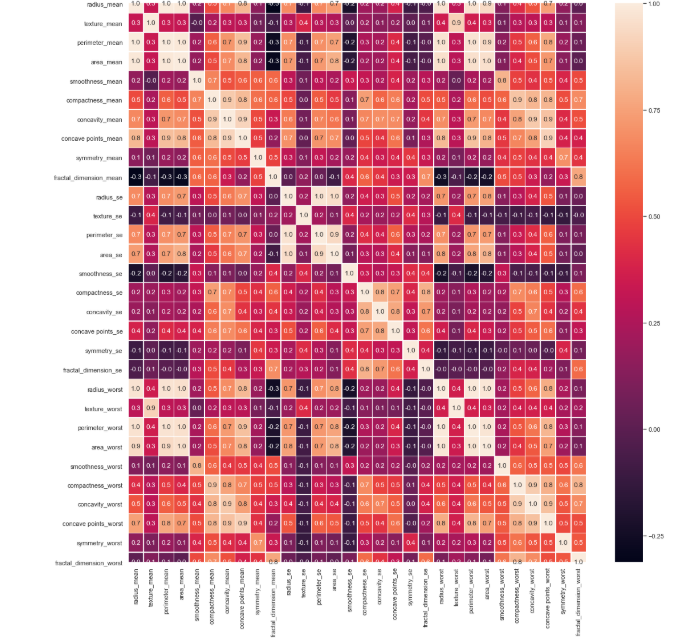
**Features of dataset are as follows:**

1. ID number
2. Diagnosis(M= malignant, B=bengin)

Ten real-valued features are computed for each cell nucleus:

1. 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(perimeter^2/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(“coastline approximation”-1)

Dataset Correlation Analysis

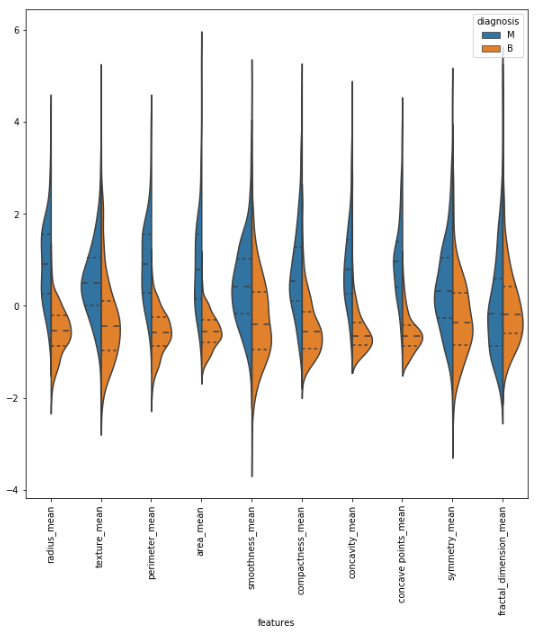


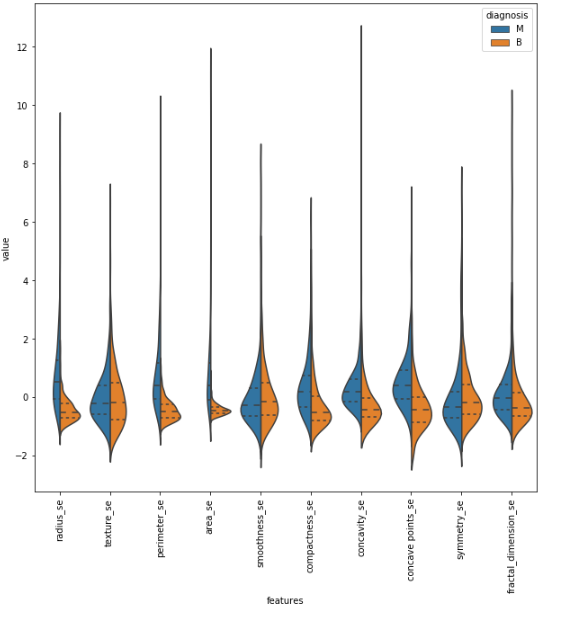
Above correlation matrix shows each feature correlation with itself as well as with other features of dataset. Its value of correlation is in between 1 to -1 and different colors of matrix shows that how much it is correlated.

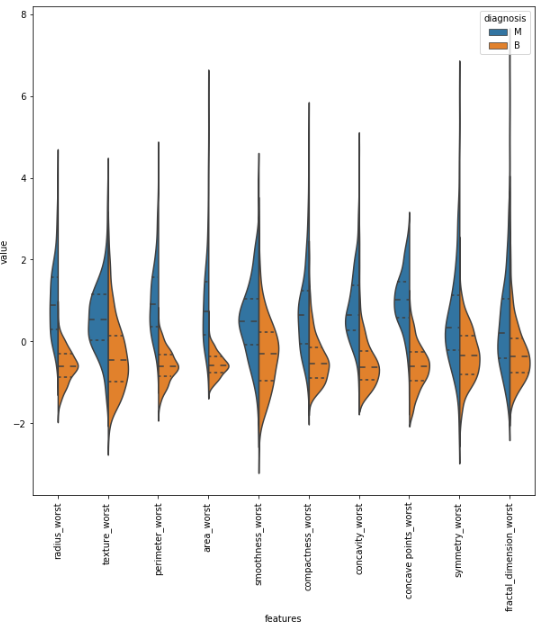
As per pre-processing purpose this is a very important step to be performed because it shows whether a feature is correlated with other feature or not, if features are correlated then it important to remove them or remove either of correlated feature.

Correlation between features should be minimum because it ultimately effect the training model and later on maybe it will lead to overfitting.

Feature Visualization







Algorithm and Techniques Used and Accuracy Achieved

**Random Forest Classifier:** Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees**.**

**Support Vector Machine:** Support-vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

**Gaussian NB:** A Gaussian Naive Bayes algorithm is a special type of NB algorithm. It’s specifically used when the features have continuous values. It’s also assumed that all the features are following a gaussian distribution i.e, normal distribution.

**Technique Used for Training Model:**

**Neural Network:** A **neural network** is trained by adjusting **neuron** input weights based on the **network's** performance on example inputs. If the **network** classifies an image correctly, weights contributing to the correct answer are increased, while other weights are decreased.

**Accuracy achieved after training model:**

**Accuracy on training data – 100%**

**Accuracy on testing data – 88%**