Breast Cancer Classification

By: Sarita Patel

Table of Contents

Executive Summary:

Breast cancer is one of the main causes of cancer death worldwide. Early diagnostics significantly increases the chances of correct treatment and survival, but this process is tedious and often leads to disagreement between pathologists. Computer-aided diagnosis systems show potential for improving the diagnostic accuracy.

Project Goals:

In this project, we compare a set of models to check which machine learning algorithm which provides the fastest and most accurate result for predicting breast cancer. So that it can help pathologists in identifying the root causes of cancer in patients.

Methodology:

In this project we first do exploratory analysis to observe different aspects of the data. We see that with the increase in the size radius of the cells, the chance of having a cancer also increases. This gave us a general idea of what kind of results we can expect. Then, the problem of multicollinearity was tackled by removing variables which seems to be like each other. After the initial analysis is done, we then reduced the number of variables to get the reduced set of variables by using principal component analysis. This provided us with 5 components which we used for our models. This analysis also gave us a brief insight about the variables which plays a vital role in predicting the cancer. We figured that the mean of concavity, area, texture and symmetry variables as well as the standard error of fractal dimension variable plays vital role in predicting the Breast Cancer.

We then compare 5 different algorithms using the components which we got from PCA to check which provides the most accurate and fast result. We repeat the same process without PCA to check whether reducing the variables was necessary or not.

Conclusion:

Decision Tree was found to be the fastest and the most accurate algorithm to detect cancer. The values with PCA and without PCA provided similar results, while the analysis using PCA provided a much faster result.

Report: Data:

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. In this problem we have used 30 different columns and we will predict the Stage of Breast Cancer M (Malignant) and B (Benign)This analysis has been done using Basic Machine Learning Algorithm with detailed explanation.

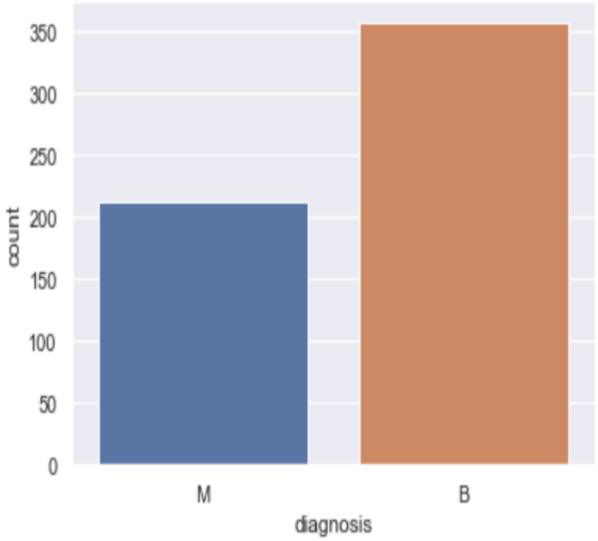
Attribute Information:

1) ID number (Patient ID number)
2) Diagnosis (Cancer type: M = malignant, B = benign)
3) Column 3-32 are divided into three parts first is Mean (3-13), Stranded Error (13-23) and Worst (23-32) and each contain 10 parameters (radius, texture, area, perimeter, smoothness, compactness, concavity, concave points, symmetry and fractal dimension)

Features:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

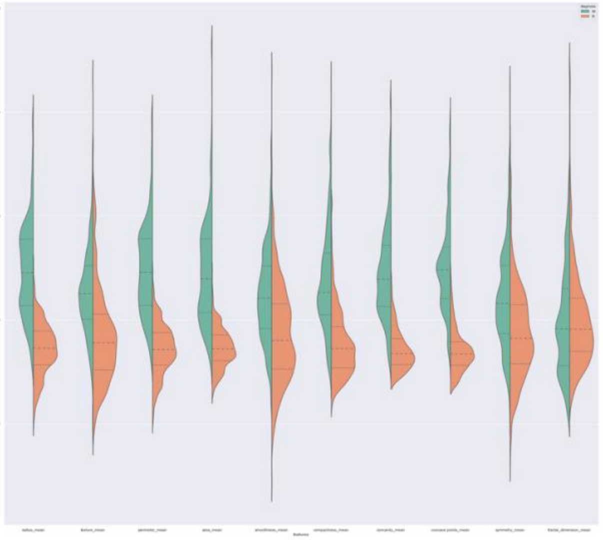
Distribution of Target variable:



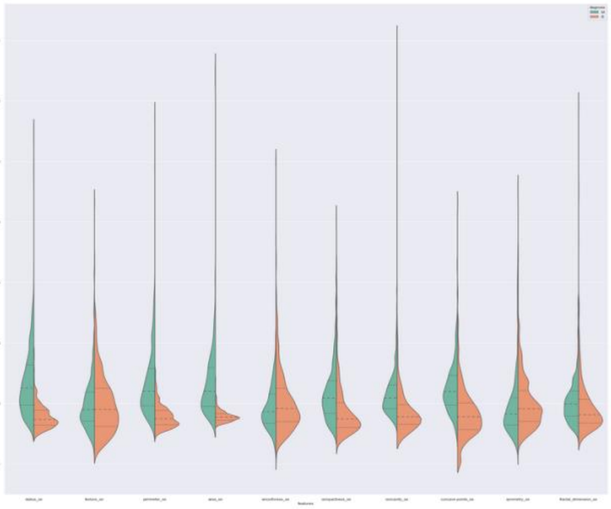
The distribution of the target variable was analyzed and seen that the number of malignant and benign cases were not equal. We can also see that they were not in the ratio of 1:2, So it is not required to upscale or downscale the target variable. We took the target variable as it is and started performing analysis.

Number of Benign: 357 Number of Malignant: 212. There are a greater number of benign cases than malignant in the dataset and the dataset looks unbiased/unbalanced but it’s not that big of a difference and it won’t impact the analysis in any ways.

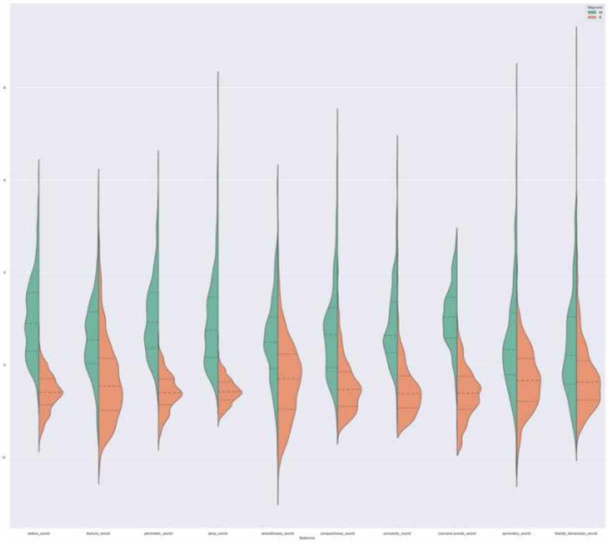
Exploratory Analysis:



The violin plot beside shows the distribution for radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, and symmetry_mean features which states that the median of the Malignant and Benign cases are well separated from each other, so these variables are pretty good for classification. However, the median of fractal_dimension_mean variable is not well separated, so it is not a good variable for classification.

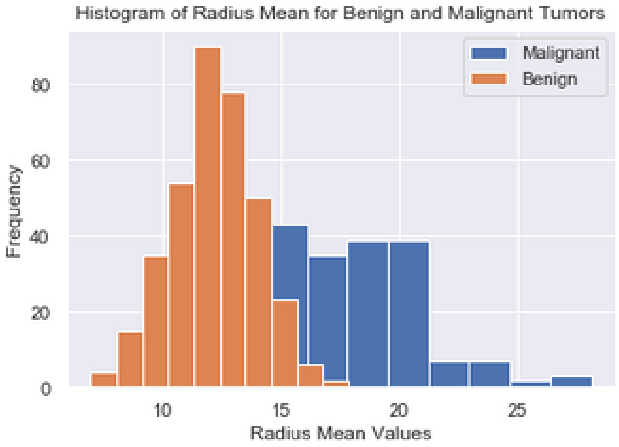


The violin plot beside shows the distribution of radius_se, perimeter_se, area_se, compactness_se, concavity_se, concave points_se features which states that the median of the Malignant and Benign cases are well separated from each other, so these variables are pretty good for classification. However, the median of texture_se, smoothness_se, symmetry_se and fractal_dimension_mean features are not well separated, so they are not a good set of variables for classification.



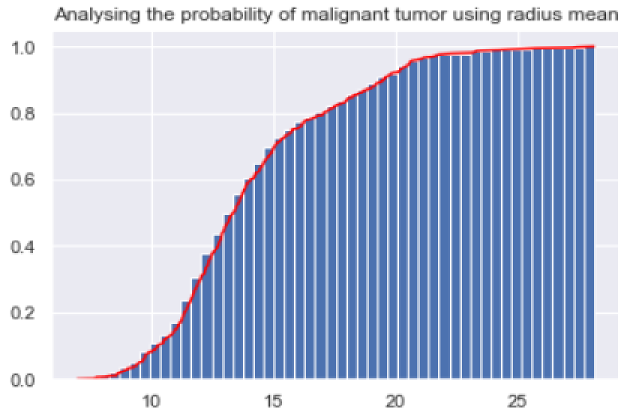
The violin plot beside shows the distribution of radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst and fractal_dimension_worst features,which states that the median of the Malignant and Benign cases are well separated from each other, so these variables are pretty good for classification

We found that these below mentioned five variables have higher values in five different PCA components so to analyse that we performed histogram, probability plot for these variables. We will analyze and explore radius_mean feature in detail to understand the impact of this variable on target variable. Let's first look at the distribution of radius_mean feature by diagnosis.



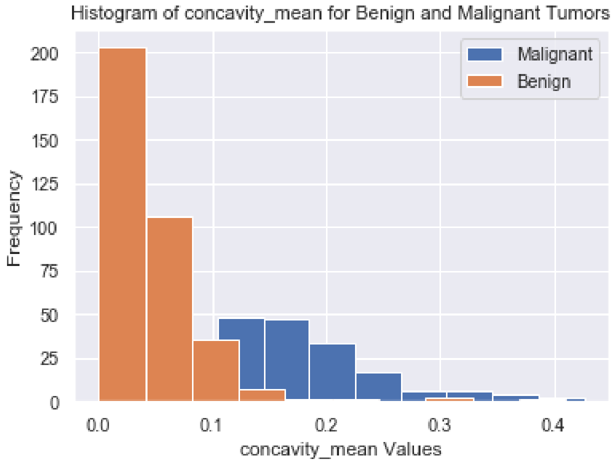
We can see that the Malignant cancer has much higher values for radius_mean when compared with Benign cancer. So, our findings were true, radius_mean is a good variable for identifying Malignant cancer.'

We further analyze the probability of malignant tumor using cases of radius_mean variable.



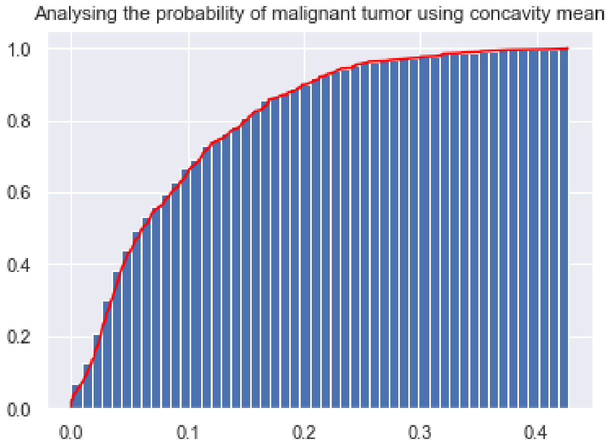
From the above plot we can analyze that for the cells with radius_mean of 20 has 85% probability of being malignant cancer. This can be useful for early stage cancer detection.

We will analyze and explore the Concavity_mean feature in detail to understand the impact of this variable on the target variable. Let's first look at the distribution of Concavity_mean feature by the diagnosis type.



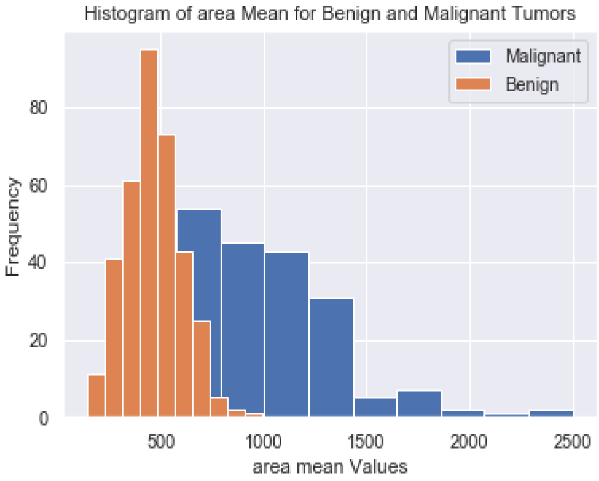
We can see that the malignant cancer has much higher values for Concavity_mean when compared with Benign cancer. So, our findings were true, Concavity_mean is a good variable for Malignant Benign cancer.

We further analyze the probability of Malignant tumor using cases of Concavity_mean variable.



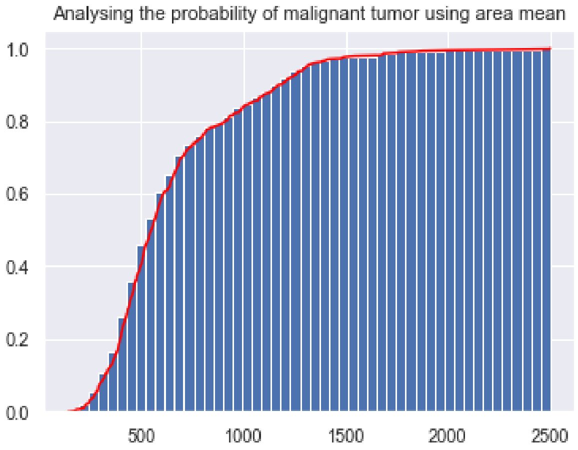
From the above plot we can analyze that for the cells with Concavity_mean of 0.4 has 100% probability of being Malignant cancer. This can be useful for early stage cancer detection.

We will analyze and explore the area_mean feature in detail to understand the impact of this variable on the target variable. Let's first look at the distribution of area_mean feature by the diagnosis type.



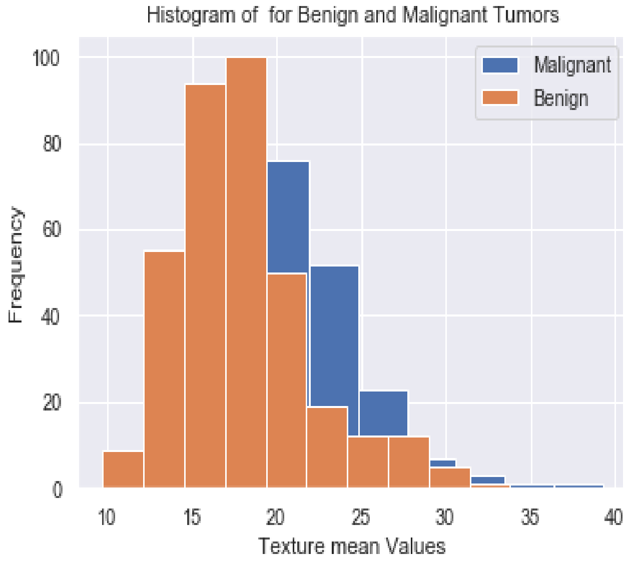
We can see that the Malignant cancer has much higher values for area_mean when compared with Benign cancer. So, our findings were true, area_mean is a good variable for identifying Malignant cancer.

We further analyze the probability of Malignant tumor using cases of area_mean variable.



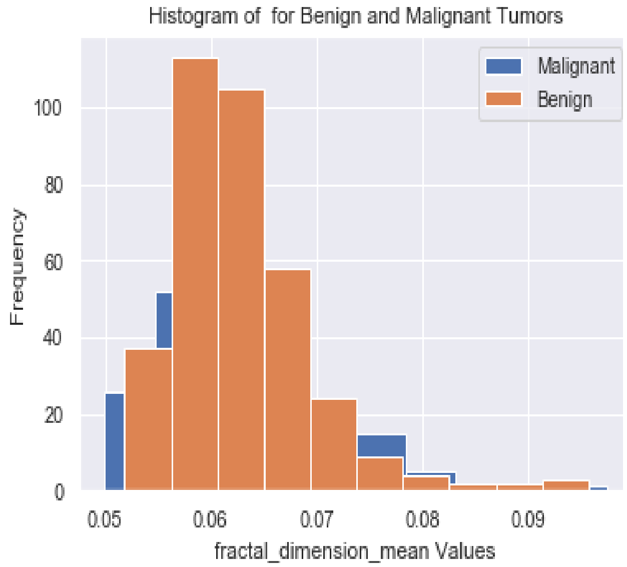
From the above plot we can analyze that for the cells with area_mean of 2000 has 100% probability of being Malignant cancer. This can be useful for early stage cancer detection.

We will analyze and explore the texture_mean feature in detail to understand the impact of this variable on the target variable.Let's first look at the distribution of texture_mean feature by the diagnosis type.



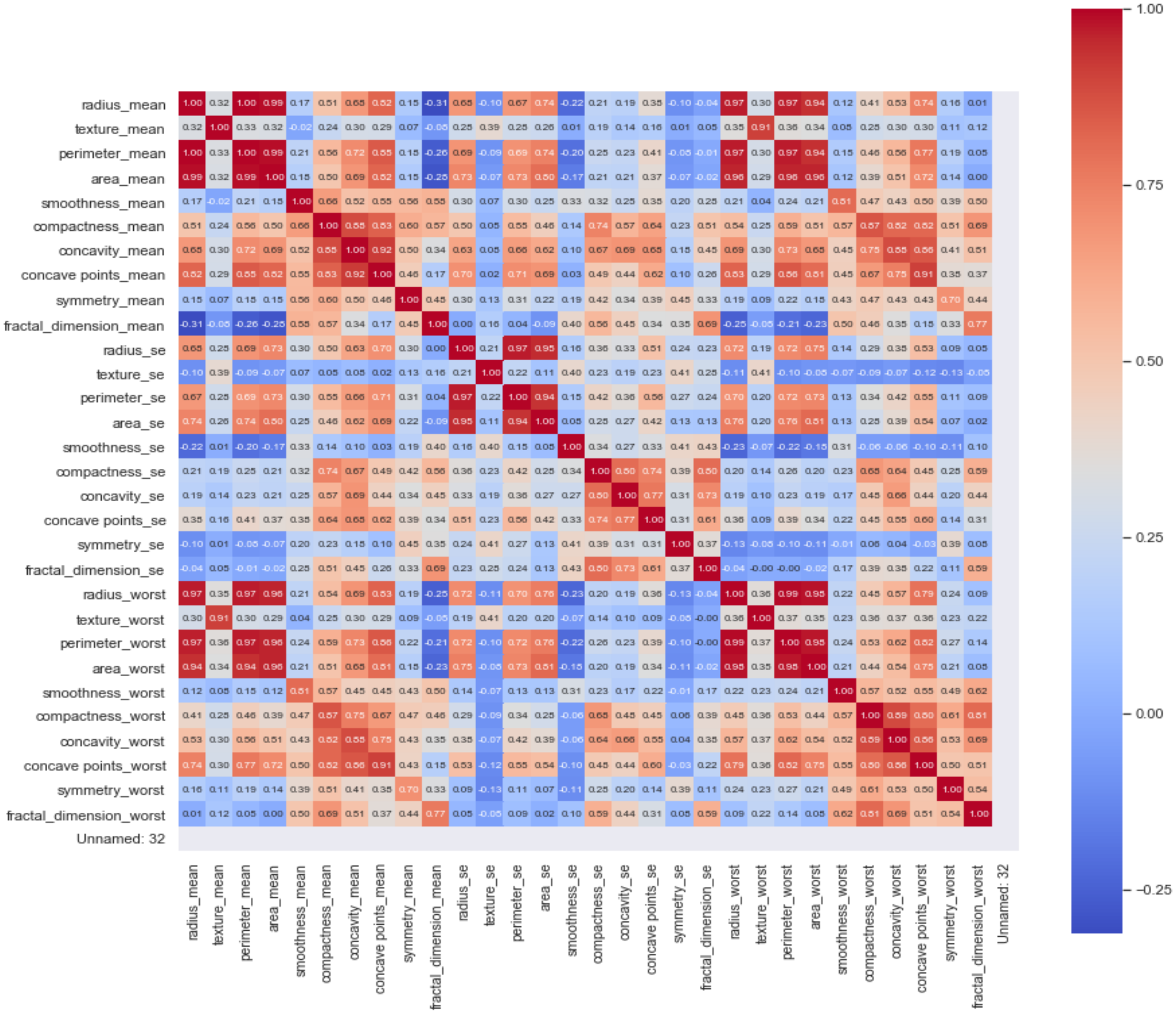
We can see that the Malignant cancer as well as Benign Cancer has much higher values for Texture_mean. So, our findings were true, Texture_mean is not a good variable for identifying cancer.

We will analyze and explore the fractal_dimension_mean feature in detail to understand the impact of this variable on the target variable. .Let's first look at the distribution of fractal_dimension_mean feature by the diagnosis type.



We can see that the Malignant cancer as well as Benign Cancer has much higher values for fractal_dimension_mean. So our findings were true, fractal_dimension_mean is not a good variable for identifying cancer.

Multicollinearity:



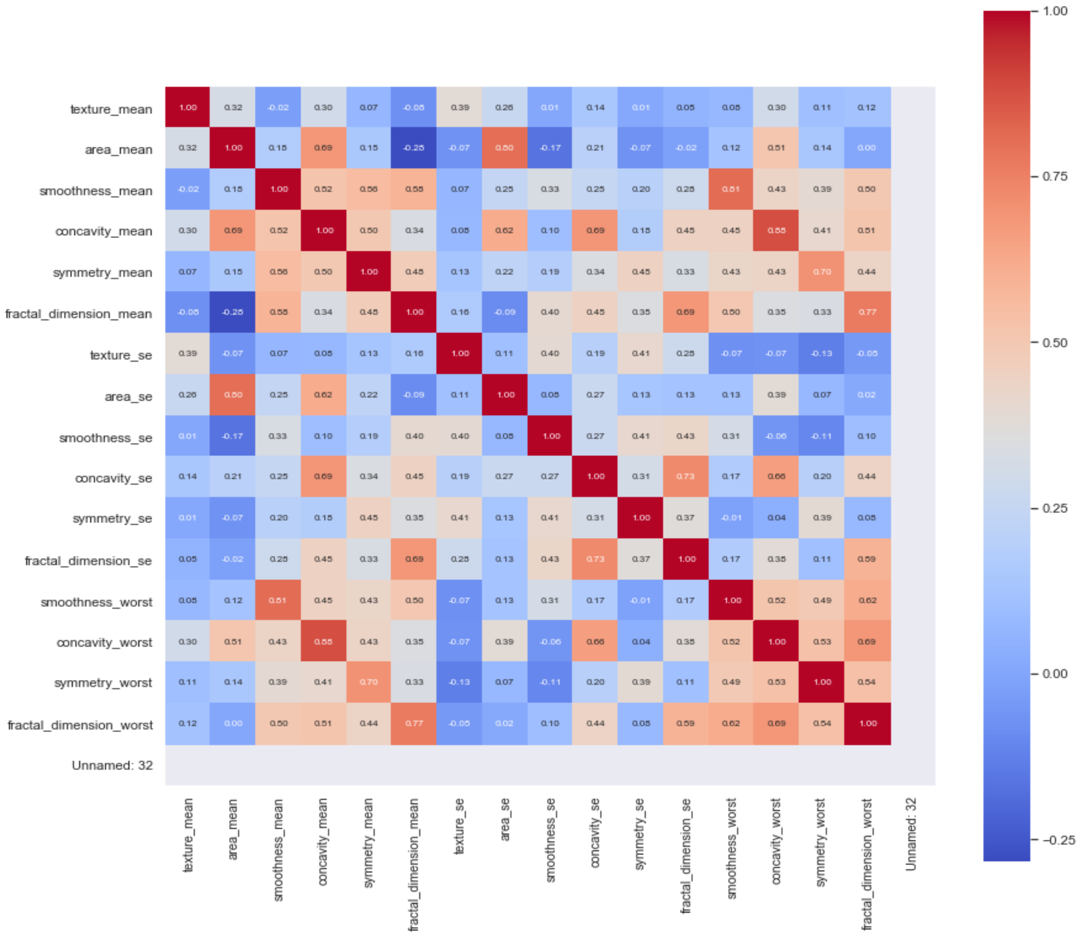
There was a high correlation in a lot of variables as seen from the heat map hence variables with high multicollinearity were removed.

Observation removed for Mean: The radius_mean, perimeter_mean, area_mean and concavepoint_mean is highly correlated as expected, so we will use anyone of them. The compactness_mean, concavity_mean and concavepoint_mean is highly correlated so we will use compactness_mean. So, selected Parameter for use is perimeter_mean, texture_mean, smoothness_mean, compactness_mean, symmetry_mean and fractal_dimension_mean.

Observation removed for standard error: The radius_se, perimeter_se, area_se and concavepoint_se are highly correlated as expected, so we will use anyone of them. The compactness_se, concavity_se and concavepoint_se are highly correlated so we will use compactness_se. So, selected Parameter for use is perimeter_se, texture_se, smoothness_se, compactness_se, symmetry_se and fractal_dimension_se.

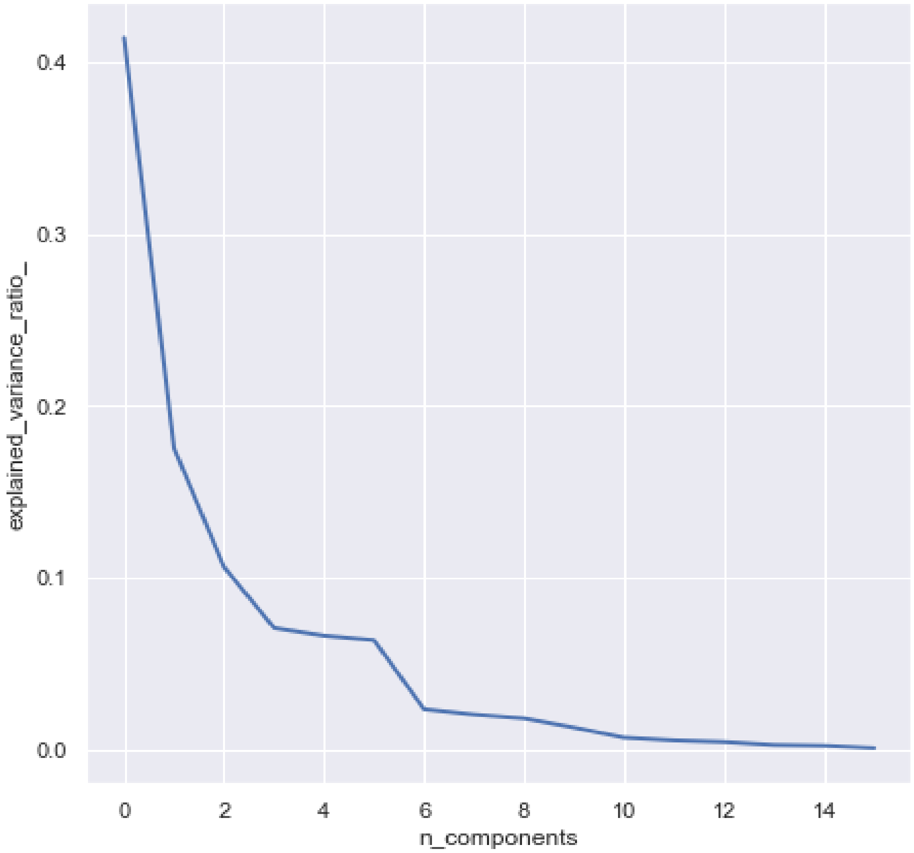
Observation removed for Worst cases: The radius_worst, perimeter_worst, area_worst and concavepoint_worst are highly correlated as expected, so we will use anyone of them. The compactness_worst, concavity_worst and concavepoint_worst are highly correlated so we will use compactness_worst. So, selected Parameter for use is perimeter_worst, texture_worst, smoothness_worst, compactness_worst, symmetry_worst and fractal_dimension_worst.

The following Pearson correlation matrix was obtained after removing highly correlated variables hence solving the problem of multicollinearity.

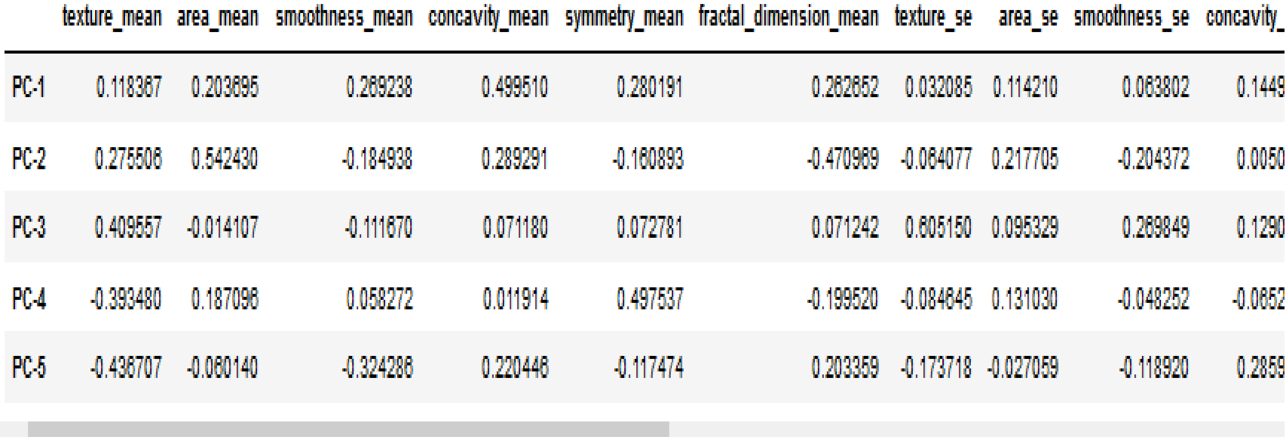


Dimensionality Reduction:

PCA was employed to reduce the number of features and to avoid over-fitting. The optimal number of principal components were discovered to be 5 which explained 84% of the variance.



These are most important Features in the Dataset which were results from the five Principal Components as this explains 85% variance of the Data.



	texture_mean	area_mean	smoothness_mean	concavity_mean	symmetry_mean	fractal_dimension_mean	texture_se	area_se	smoothness_se	concavity_
PC-1	0.118367	0.203695	0.269238	0.499510	0.280191	0.262652	0.032085	0.114210	0.063802	0.1449
PC-2	0.275506	0.542430	-0.184938	0.289291	-0.160893	-0.470969	-0.064077	0.217705	-0.204372	0.0050
PC-3	0.409557	-0.014107	-0.111670	0.071180	0.072781	0.071242	0.605150	0.095329	0.269849	0.1290
PC-4	-0.393480	0.187096	0.058272	0.011914	0.497537	-0.199520	-0.084645	0.131030	-0.048252	-0.0652
PC-5	-0.436707	-0.060140	-0.324286	0.220446	-0.117474	0.203359	-0.173718	-0.027059	-0.118920	0.2859

Methodology:
We applied the following machine learning techniques to predict if a patient has cancer:

Logistic Regression
KNN
Decision Tree
Random Forest
SVM

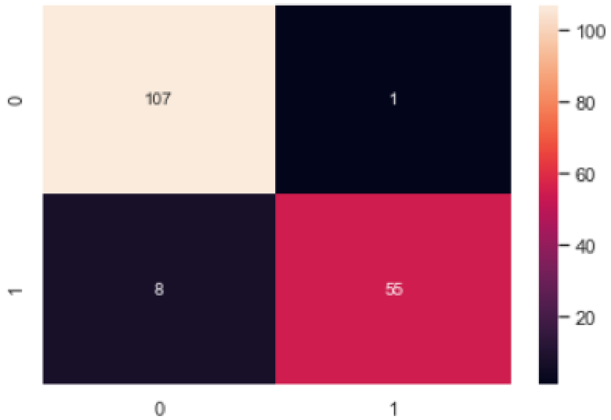
Every model was run with the principal components and with the dataset variables as parameters to compare the efficiency of the models.

Models:

Logistic Regression:
Logistic regression is a statistical model that uses a logistic function to model a binary dependent variable.

Results for Logistic Regression:
Logistic Regression Training Dataset Accuracy: 93%
Logistic Regression Test Dataset Accuracy: 94%
Run time: 0.034 Seconds.

Confusion Matrix:



Explanation:
There were 107 true negative and 8 false negative observations and 56 true positive and 1 false positive observation.
Precision value for malignant cancer is 98% which indicates that 98% of the cancer cases are predicted correctly.

KNN:

KNN is a supervised machine learning algorithm Which relies on labeled input data to learn a function that produces an appropriate output on the given new unlabeled data.

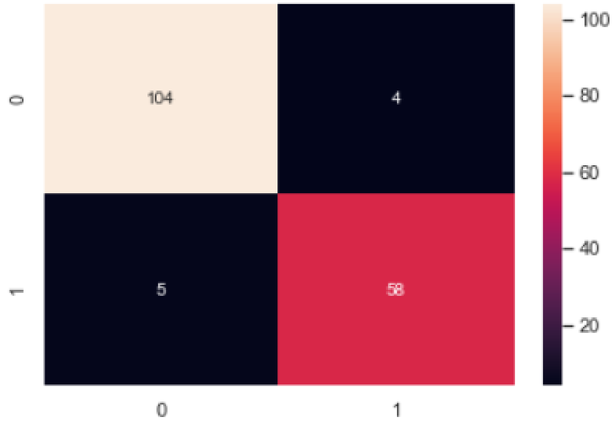
Result for KNN:

KNN classifier Training Dataset Accuracy: 100%

KNN classifier Test Dataset Accuracy: 94%

Run time: 0.020 Seconds

Confusion Matrix:



Explanation:

There were 104 true negative and 5 false negative observations and 56 true positive and 4 false positive observations.

Precision value for malignant cancer is 94% which indicates that 94% of the cancer cases are predicted correctly.

Decision Tree:

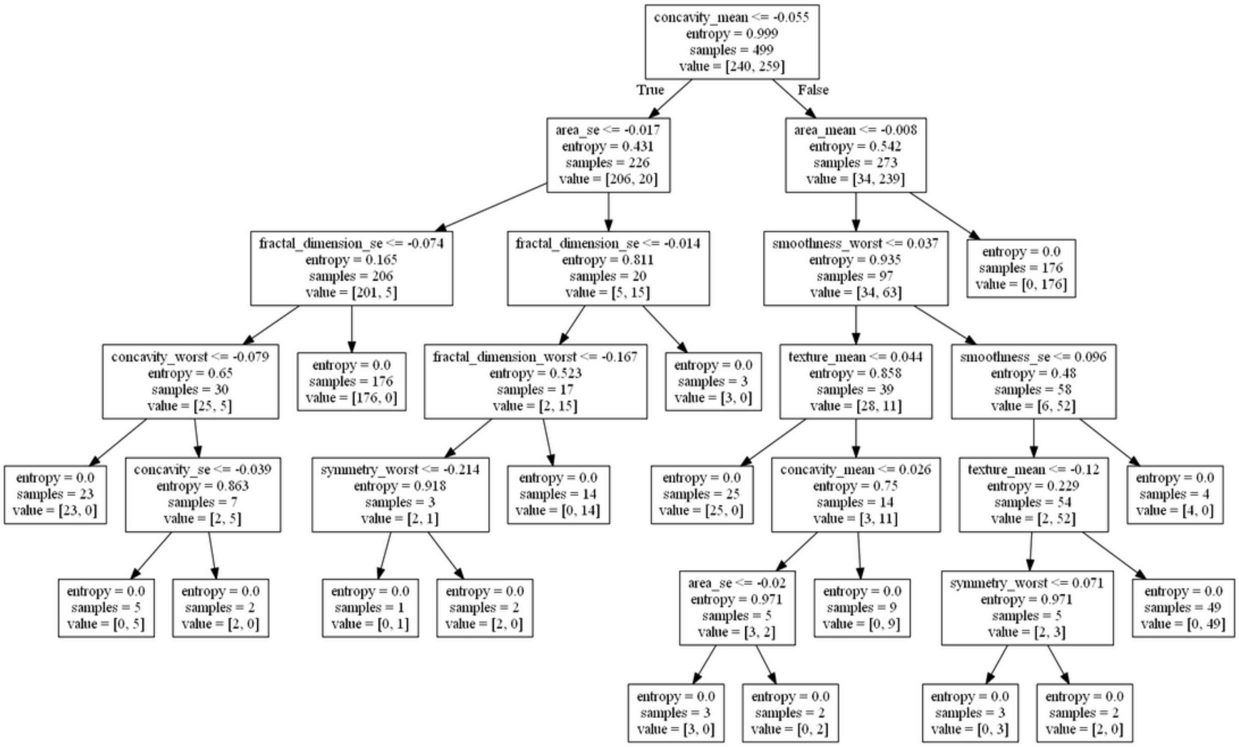
Decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Result for Decision Tree:

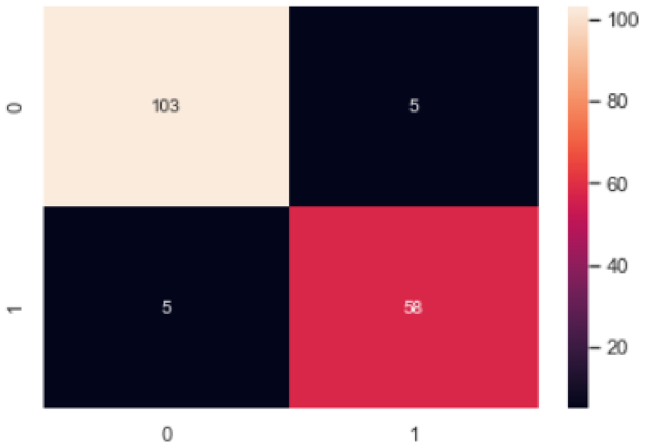
Decision Tree classifier Training Dataset Accuracy: 94%

Decision Tree classifier Test Dataset Accuracy: 99%

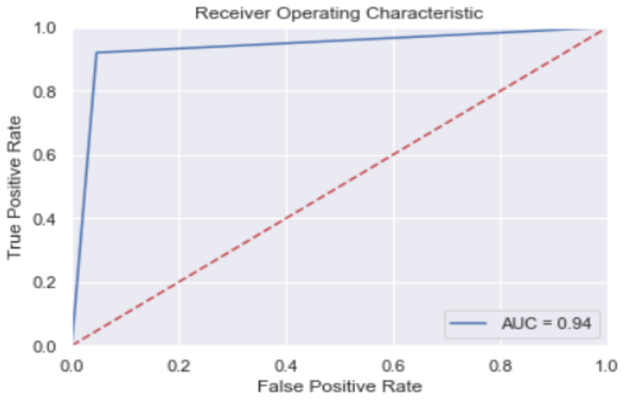
Run time: 0.013Seconds



Confusion Matrix:



ROC for Decision Tree:



Explanation:

There were 103 true negative and 5 false negative observations and 56 true positive and 5 false positive observations.

Precision value for malignant cancer is 92% which indicates that 92% of the cancer cases are predicted correctly.

Random Forest:

Random forests are an ensemble learning method for classification and regression that operates 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.

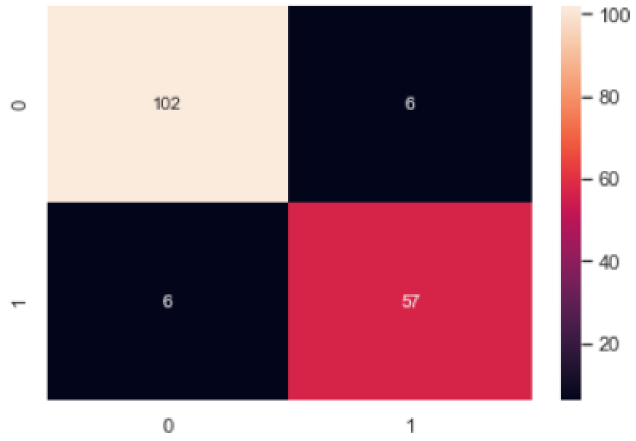
Result for Random Forest:

Random forest Training Dataset Accuracy: 98%

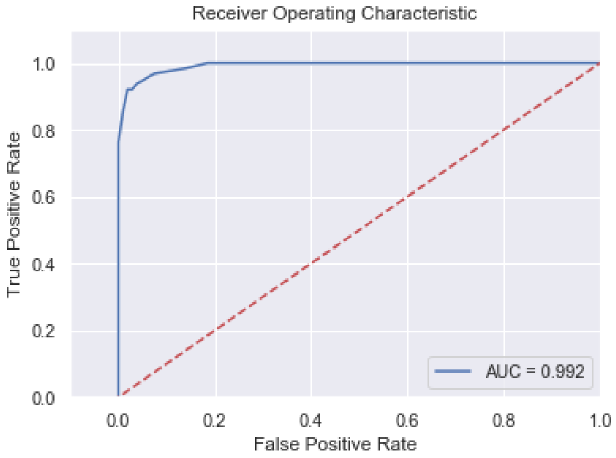
Random forest Test Dataset Accuracy: 92%

Run time: 0.011Seconds

Confusion Matrix:



ROC for Random Forest:



Explanation:

There were 102 true negative and 6 false negative observations and 57 true positive and 6 false positive observations.

Precision value for malignant cancer is 90% which indicates that 90% of the cancer cases are predicted correctly.

SVM:

A Support Vector Machine is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (*supervised learning*), the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

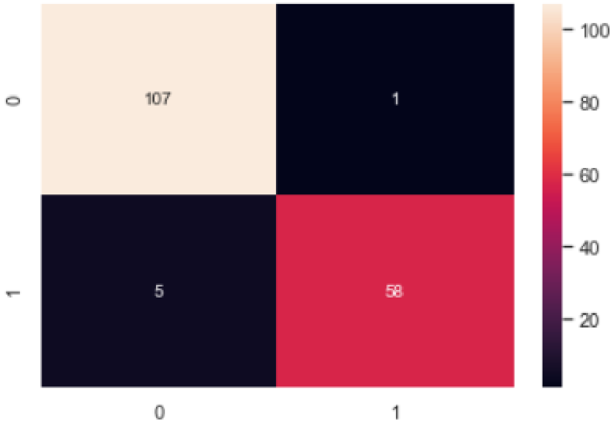
Result for SVM:

support vector machine classifier train Dataset Accuracy: 95%

support vector machine classifier Test Dataset Accuracy: 96%

Run Time: 0.010sec

Confusion Matrix:



Explanation:

There were 107 true negative and 5 false negative observations and 56 true positive and 1 false positive observations.

Precision value for malignant cancer is 98% which indicates that 98% of the cancer cases are predicted correctly.

Model Comparison with PCA:

	Training	Testing	Computation Time
Logistic Regression	93.9%	94.7%	0.034s
KNN	100%	94.1%	0.020s
Decision Tree	94.1%	99.7%	0.013s
Random Forest	98.9%	92.9%	0.011s
SVM	95.7%	96.5%	0.010s

Model Comparison without PCA:

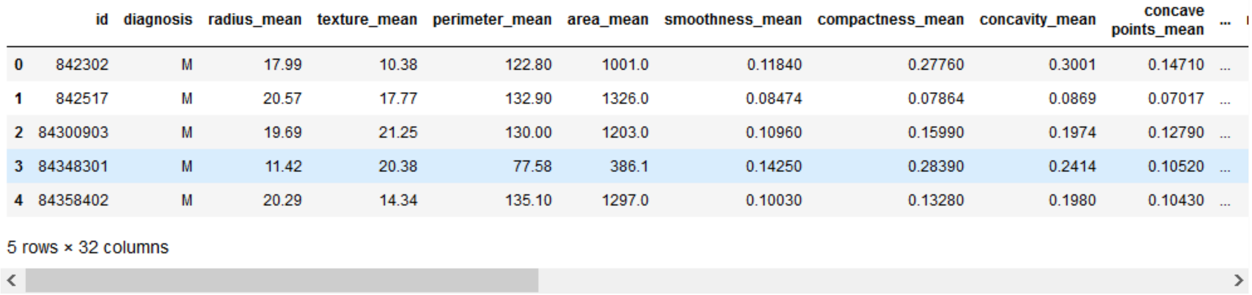
	Training	Testing	Computation Time
Logistic Regression	93.9%	94.7%	0.031s
KNN	100%	94.1%	0.044s
Decision Tree	93.5%	100%	0.063s
Random Forest	99.4%	95.3%	0.028s
SVM	96.2%	96.4%	0.011s

The run time for models using PCA was seen to be lower than without using PCA while the accuracy was seen to be similar hence performing dimensionality reduction provides a better model.

Conclusion:

Decision Trees was found to be the most accurate model with a test accuracy of 99.7%. It also provided the fastest result performing the analysis on the dataset in 0.013s while logistic regression was the slowest (0.034s) algorithm and Random Forest provides the least accurate result with only 92.9% accuracy on the test data. Hence Decision Tree is the best model to predict breast cancer.

Data Sample:



	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...

5 rows × 32 columns

Documented Code:

Refer Machine_learning_FinalProject.ipynb Notebook

Reference:

https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
https://seaborn.pydata.org/generated/seaborn.violinplot.html