MOTIVATION:

Cancer is a disease that has become rather rampant in the past decade and breast cancer is the second most common cancer in women.

Only in the United states, 1 out of 8 women is said to be diagnosed with breast cancer, and so far in the year 2017 alone, around 46000 women have succumbed to breast cancer

Death rates have significantly decreased. However, it is still a matter of concern.

INTRODUCTION:

Three methods are generally used to diagnose a patient, mammography, FNA and surgical biopsy.

Out of the three methods, surgical biopsy has the highest sensitivity (approximately 100% accuracy) compared to the sensitivity of mammography (63% to 97% accuracy) and FNA (Fine Needle Aspiration) with visual Interpretation (65% to 98%).

DETAILS ABOUT THE DATASET:

The dataset that we used was obtained from the UCI machine learning repository.

The dataset was created in the year 1995, by Dr. William Holdberg , Mr. Nick Street and Mr. Olvi L Mangasarian.

Number of samples: 569

Total number of attributes: 32

A digitized image of the Fine Needle Analysis is used to calculate the following 10 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)
```

The mean, standard error and the worst values were calculated for each of the above-mentioned features, thus making the total number of attributes as 30.

The response variable for our dataset is Diagnosis, which has two possible outputs: **Malignant and benign.**

In our dataset out of the 569 instances we have 357 cases which are benign and 212 cases which are malignant.

Benign tumour: A tumour is said to be benign, when it does not invade the surrounding organs, but can be quite dangerous when it develops in vital body parts.

Malignant tumour: A malignant tumour is one that spreads to the organs surrounding it rapidly.

GOAL:

In our project, we would like to apply various machine learning models to and compare the models based on the various accuracy metrics.

LITERATURE SURVEY:

The accuracy of predicting a tumour varies with the radiologists' experience. This can be improved with computer aid. Researchers have been working on building decision systems, which will help reduce the misdiagnoses.

Floyd et al developed an Artificial neural network to predict the malignancy of tumours, which proved more efficient than the predictions made by radiologists.

They recorded a relative sensitivity of 1.0 and a specificity of 0.59

More recent studies incorporated recent data mining techniques such as SVM, Naive Bayes and CART models, Aruna et al, compared the performance of the above-mentioned models and concluded that SVM with a radial basis kernel outperformed the other models.

They observed an accuracy of 98.06% for the Support vector machine model.

Another research paper on the same by Diaz-Uriarte and De Andres concluded that Random forest had the best accuracy when compared to other classification models they implemented.

A similar research was done on breast cancer survivability by Delen D, walker G and Kadam A, where they implemented three data mining tehcniques: ANN, Decision trees and Logistic regression, out of the three methods mentioned, they concluded that Decision tree had the best prediction with 93.6 % accuracy, followed by ANN with 91.2% followed by Logistic regression with 89.2%

Although breast cancer specific related research is not abundantly available for recent Machine learning models, their implementation in other clinical health system research can be used as a framework to utilise it in breast cancer research.

EXPLORATORY ANALYSIS:

Structure of the Dataset:

```
Here's the dimensions of our data frame:
 (569, 31)
Here's the data types of our columns:
 diagnosis                 object
radius_mean               float64
texture_mean              float64
perimeter_mean            float64
area_mean                 float64
smoothness_mean           float64
compactness_mean          float64
concavity_mean            float64
concave_points_mean       float64
symmetry_mean             float64
fractal_dimension_mean    float64
radius_se                 float64
texture_se                float64
perimeter_se              float64
area_se                   float64
smoothness_se             float64
compactness_se            float64
concavity_se              float64
concave_points_se         float64
symmetry_se               float64
fractal_dimension_se      float64
radius_worst              float64
texture_worst             float64
perimeter_worst           float64
area_worst                float64
smoothness_worst          float64
compactness_worst         float64
concavity_worst           float64
concave_points_worst      float64
symmetry_worst            float64
fractal_dimension_worst   float64
```

We can observe that, we have a total of 569 records and a total of 31 features, including the output variable.

All the features are of float type and the response variable diagnosis is of object type.

Distribution of the output variable:

```
0    357
1    212
Name: diagnosis
```

0: Benign tumour

1: Malignant tumour

The percentage of Malignant Diagnoses: 37.258%

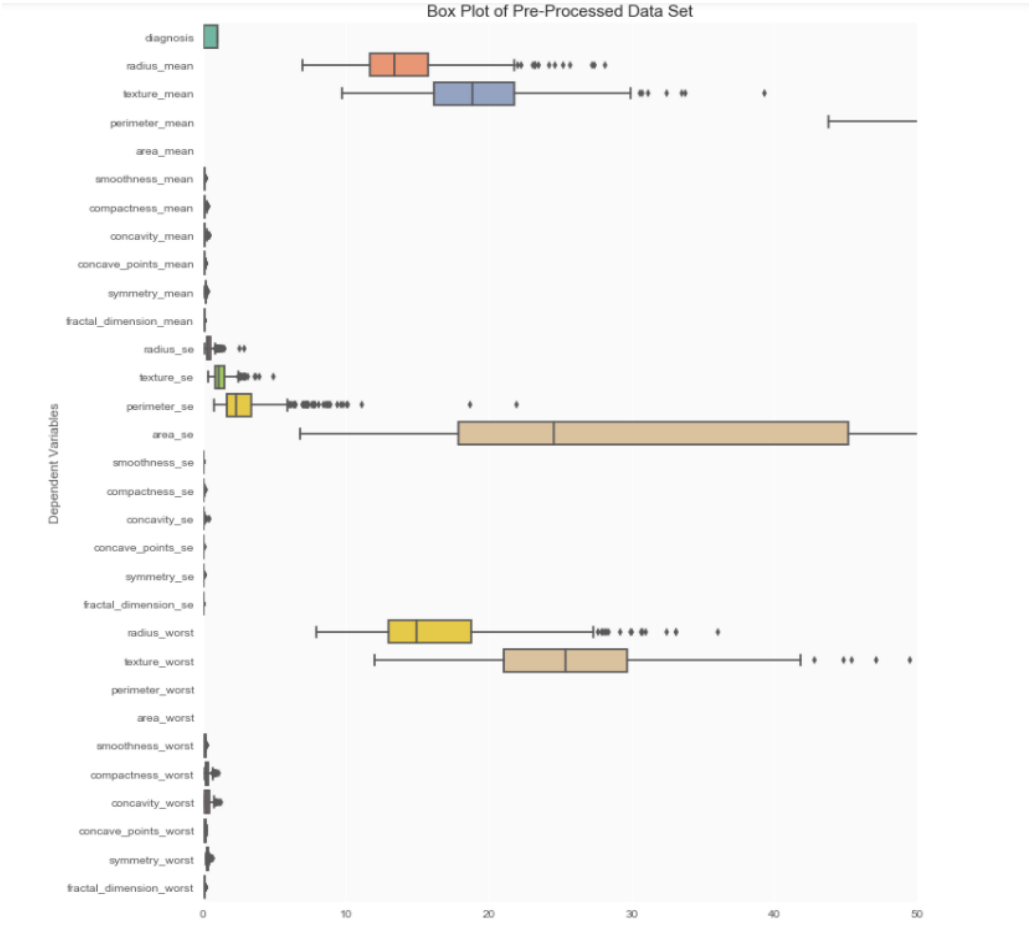
The percentage of Benign Diagnoses: 62.742%

We can see that our dataset does not suffer from class imbalances.

Out[10]:

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave_points_mean	sy
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	
mean	0.372583	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	
std	0.483918	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	
min	0.000000	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	
25%	0.000000	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	
50%	0.000000	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	
75%	1.000000	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	
max	1.000000	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	

The above result shows the basic statistics of each variable. By looking at the mean values, we can see that distributions of different variables have high variances. Some variables have small mean values, and some have higher mean values. So, we will have to normalize the data.



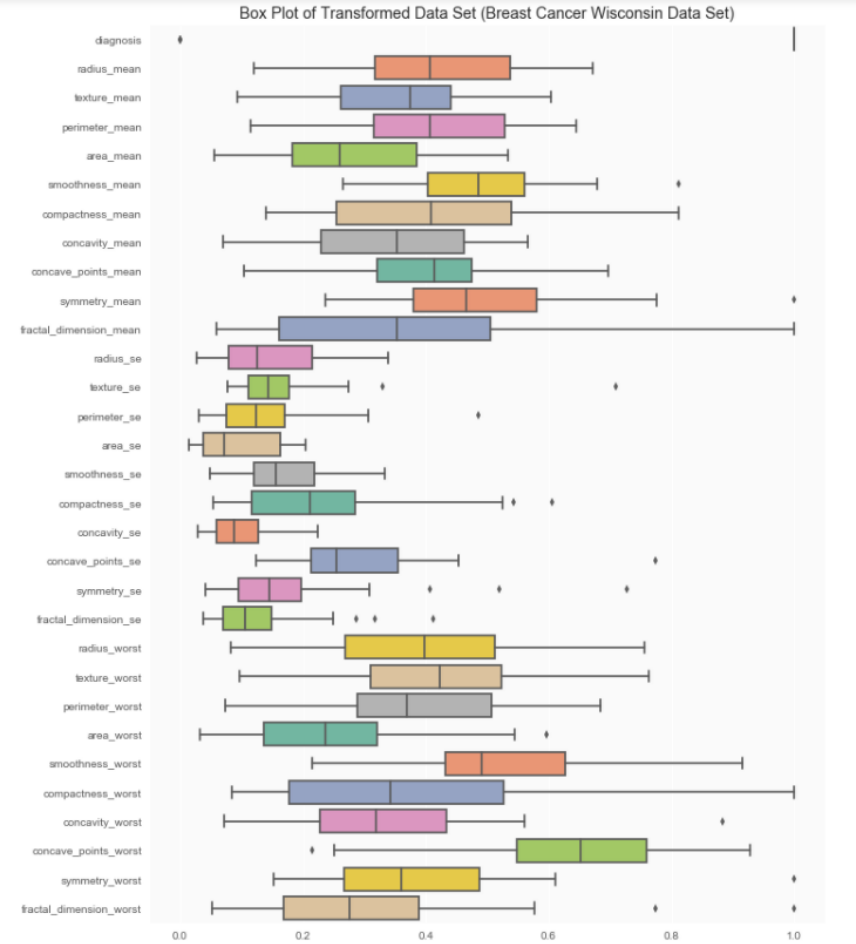
From the above box plot, we can notice the high variances in the distribution of the variables.

Out[15]:

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave_points_mean	syn
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	
mean	0.372583	0.338222	0.323965	0.332935	0.216920	0.394785	0.260601	0.208058	0.243137	
std	0.483918	0.166787	0.145453	0.167915	0.149274	0.126967	0.161992	0.186785	0.192857	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.223342	0.218465	0.216847	0.117413	0.304595	0.139685	0.069260	0.100944	
50%	0.000000	0.302381	0.308759	0.293345	0.172895	0.390358	0.224679	0.144189	0.166501	
75%	1.000000	0.416442	0.408860	0.416765	0.271135	0.475490	0.340531	0.306232	0.367793	
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

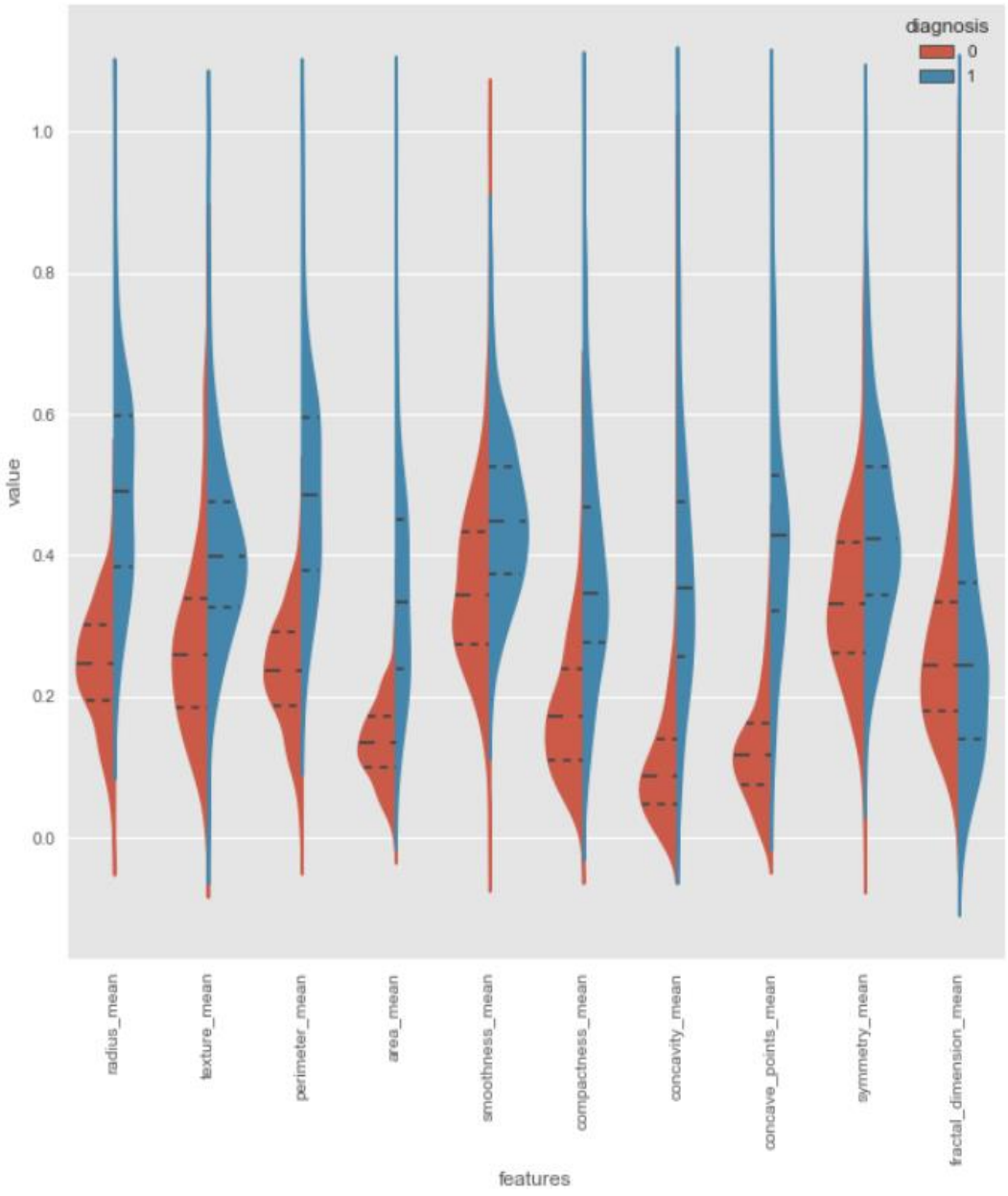
The above result shows the basic statistics of each variable after scaling.

Box Plot of all the variables (after scaling):

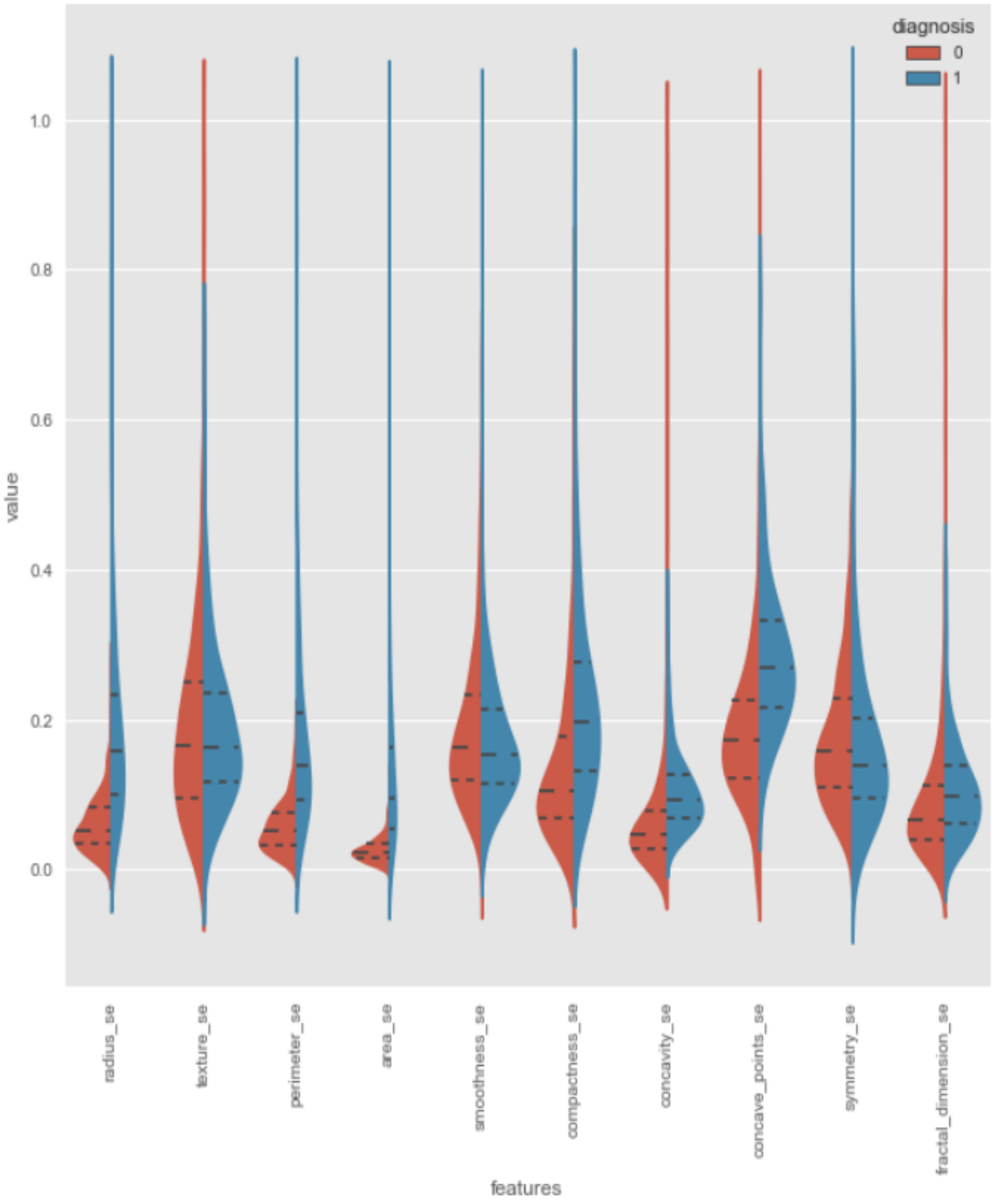


We can clearly see the differences from the first box plot.

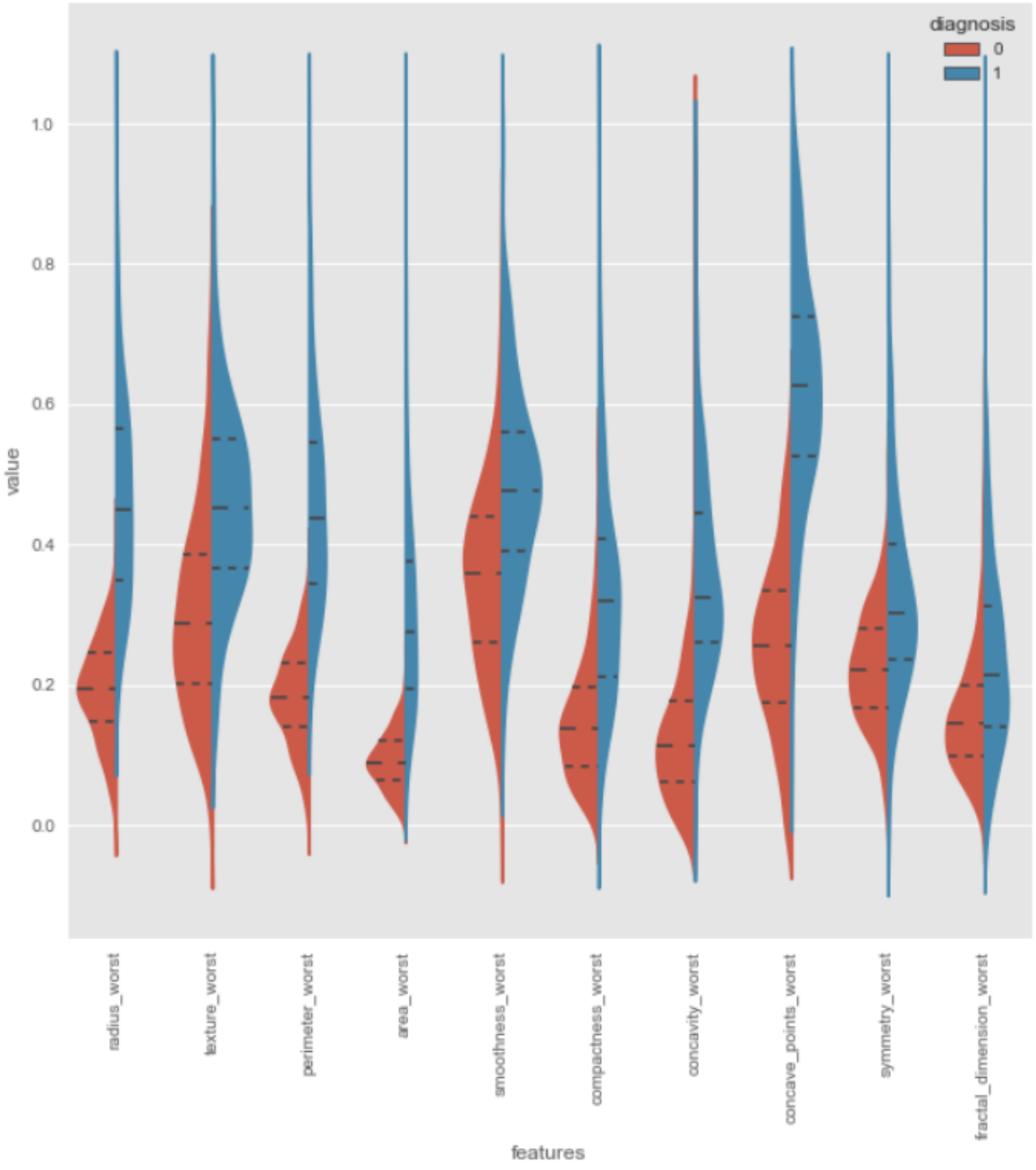
Violin Plots for the first 10 variables:



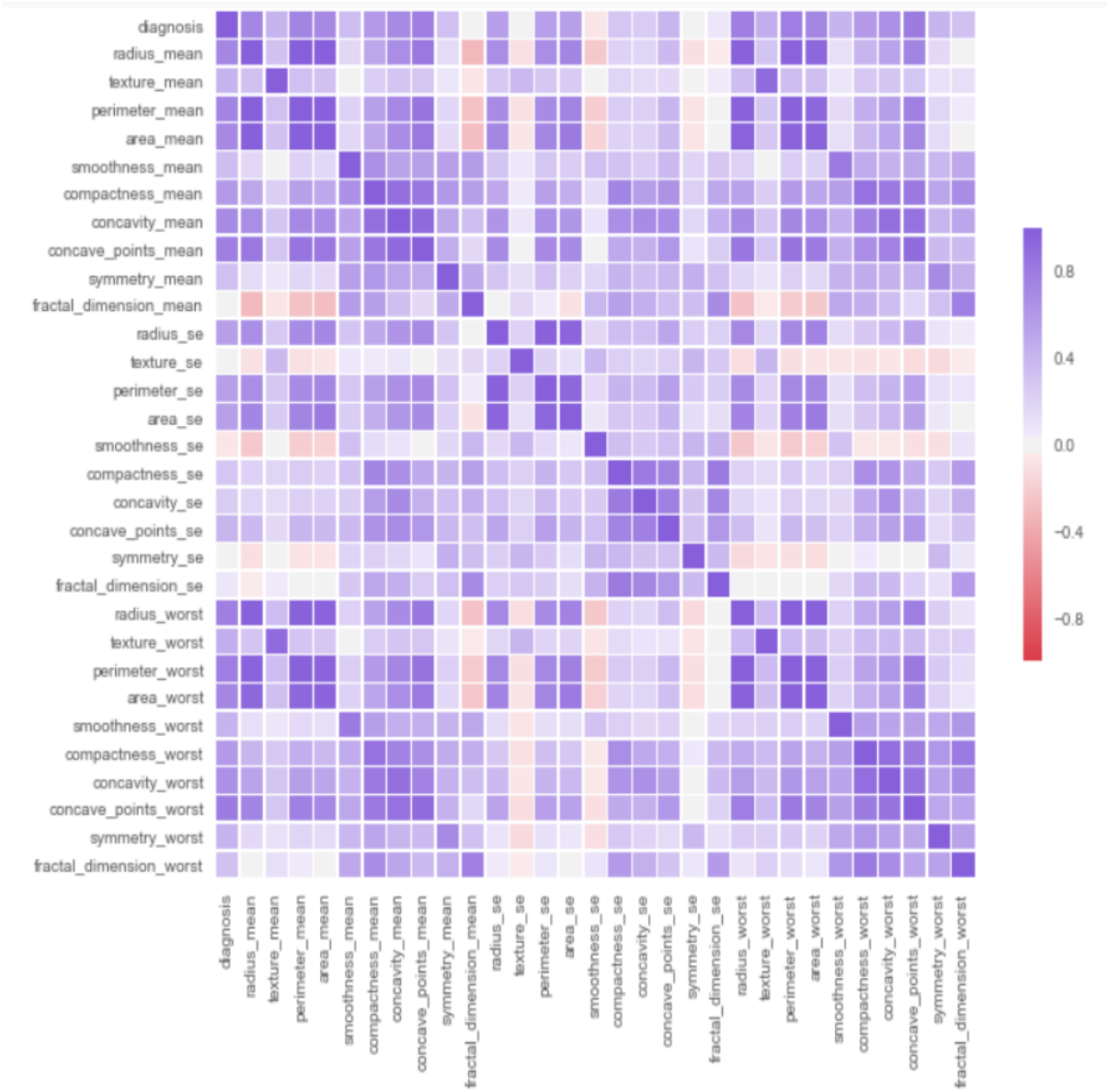
Violin Plots for the variables 11-20:



Violin Plots for the variables 21-30:



PEARSON CORRELATION MATRIX:



We can see that most of the variables have positive correlation.

FEATURE SELECTION:

From the correlation matrix, we can see that **radius_mean, perimeter_mean and area_mean** are correlated with each other. If we look at the violin graphs, we cannot make exact separation between the radius_mean and perimeter_mean. So, I have excluded them and included only area_mean.

Similarly, **area_worst and area_mean** are correlated, I have included **area_mean**. **texture_mean and texture_worst are correlated,** I have included **texture_mean. compactness_mean, concavity_mean and concave points_mean** are correlated with each other, I have included **concavity_mean**. **radius_se, perimeter_se and area_se** are correlated with each other, I have included **area_se**. **radius_worst, perimeter_worst and area_worst** are correlated with each other, I have included **area_worst. compactness_worst, concavity_worst and concave points_worst** are correlated with each other, I have included **concavity_worst**. **compactness_se, concavity_se and concave points_se** are correlated with other, I ha **concavity_se**.

MODEL IMPLEMENTATIONS:

We have split the dataset in to training and test sets.

Training Dataset: 80% of the entire dataset.

Testing Dataset: 20% of the entire dataset.

We have implemented the following models on our dataset:

a) **K-Nearest Neighbors**

b) **Gaussian Naive Bayes**

c) **AdaBoost**

d) **Random Forest**

e) **Linear SVC**

EVALUATION METRICS:

Accuracy: It is measured as number of correct predictions divided by total number of the dataset.

Accuracy = (True Positive + True Negative) / (Total number of a dataset)

Error Rate: It is measured as number of incorrect predictions divided by total number of the dataset.

Error Rate = (False Positive + False Negative) / (Total number of a dataset)

Receiver Operating Characteristics: It is a plot between Sensitivity and (1-Specificity). It is one of the widely used measures for evaluating classifier performance. The model performs well if the area under the ROC curve is higher.

F1_Score: It is one of the measures of the test accuracy. It is the harmonic mean of the precision and recall. Precision is the number of correct positive results divided by the number of all positive results and Recall is the number of correct positive results divided by the number of positive results that should have been returned.

Cross Validation Score:

CONCLUSIONS:

MODEL	TRAINING SET EVALUATIONS			TEST SET EVALUATIONS			
	Accuracy	Error Rate	F1_score	Accuracy	Error Rate	F1_score	Cross Validation score
K-Nearest Neighbors	0.8857	0.1143	0.8301	0.939	0.061	0.9114	0.94 (+/- 0.04)
Gaussian Naïve Bayes	0.9978	0.0022	0.997	0.9386	0.0614	0.9157	0.96 (+/- 0.4)
AdaBoost	1	0	1	0.9474	0.0526	0.9318	0.93 (+/- 0.03)
Random Forest	0.9934	0.0066	0.991	0.974	0.026	0.9647	0.94 (+/- 0.03)
Linear SVC	0.7407	0.2593	0.7035	0.8246	0.1754	0.82	0.77 (+/- 0.08)

By looking at all the accuracy metrics, we conclude that Random Forest performs better when compared to other models.

We should also consider False Negatives when choosing a model. Because predicting a person to not have cancer when he actually have cancer is life threatening.

REFERENCES:

- Aruna, S., Rajagopalan, S. P., & Nandakishore, L. V. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. Computer Science & Information Technology, 2, 37-45.
- Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127.
- Floyd, C. E., Lo, J. Y., Yun, A. J., Sullivan, D. C., & Kornguth, P. J. (1994). Prediction of breast cancer malignancy using an artificial neural network. Cancer, 74(11), 2944-2948
- Application of Data Mining Techniques in Improving Breast Cancer Diagnosis Josephine S. Akosa, Oklahoma State University; Shannon Kelly, Oklahoma State University