Breast Cancer Detection

Mentor

Prof. Dr. David Belanger

Submitted by,

Aishwarya Sangu, Aashitha Koushik and Rana Putta

Introduction

Objective

Data Exploration

Data Visualization

Dimensionality Reduction

Feature Scaling

Modeling

Analysis

Conclusion

Introduction

What is Breast Cancer?

A disease in which cells in the breast grow out of control.

Why is it an important topic?

1 in 8 chance a women will develop breast cancer in her lifetime.

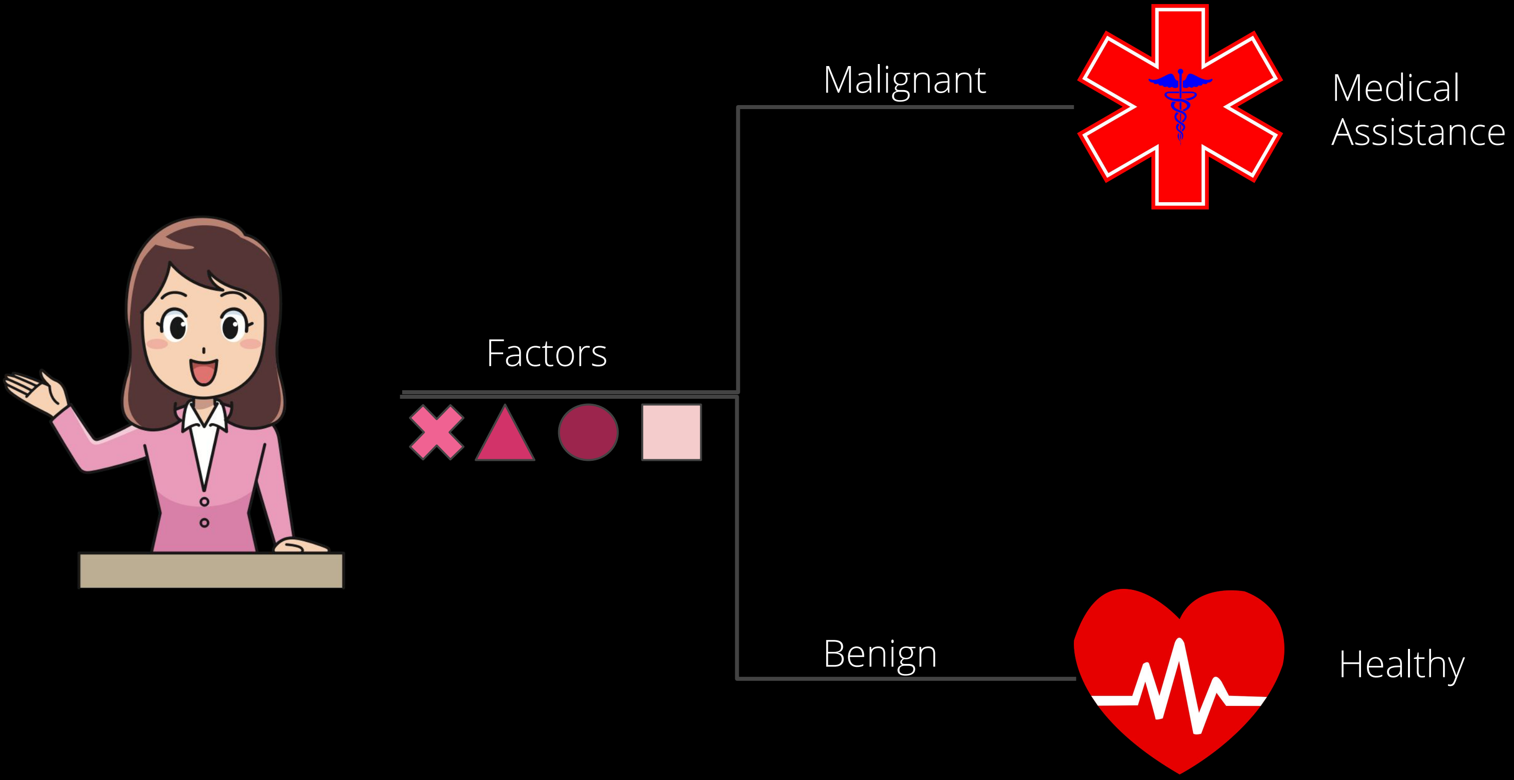
About 42,170 women will die from breast cancer in 2020.

How is it detected?

Screening, Mammogram and self-exam.

*AMERICAN CANCER SOCIETY

Objective



Factors

✖ ▲ ● ■

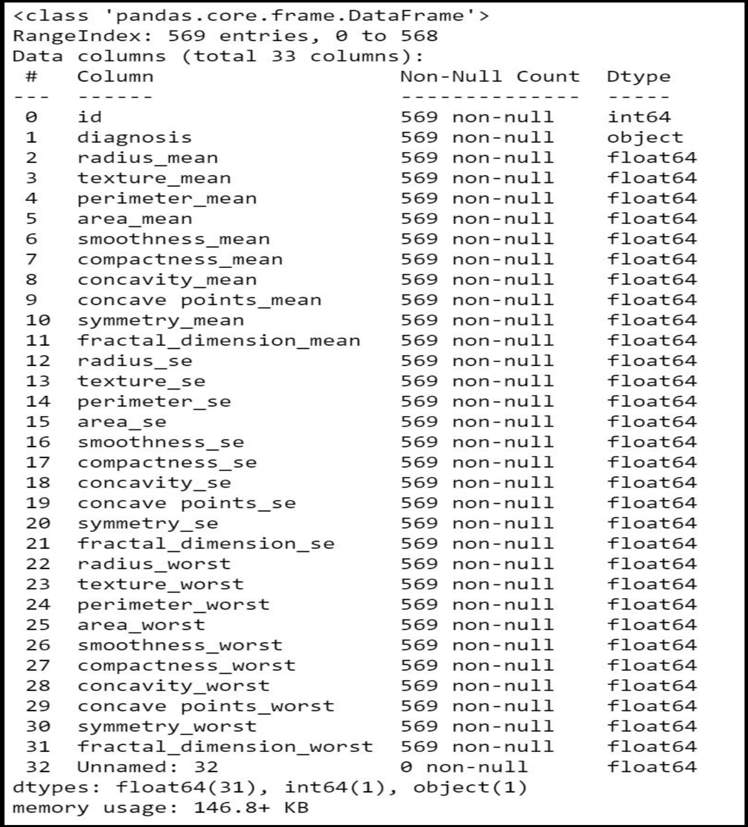
Malignant

Medical Assistance

Benign

Healthy

Data Exploration



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 33 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       569 non-null    int64
 1   diagnosis                569 non-null    object
 2   radius_mean              569 non-null    float64
 3   texture_mean             569 non-null    float64
 4   perimeter_mean           569 non-null    float64
 5   area_mean                569 non-null    float64
 6   smoothness_mean          569 non-null    float64
 7   compactness_mean         569 non-null    float64
 8   concavity_mean           569 non-null    float64
 9   concave points_mean      569 non-null    float64
 10  symmetry_mean            569 non-null    float64
 11  fractal_dimension_mean   569 non-null    float64
 12  radius_se                569 non-null    float64
 13  texture_se               569 non-null    float64
 14  perimeter_se             569 non-null    float64
 15  area_se                  569 non-null    float64
 16  smoothness_se            569 non-null    float64
 17  compactness_se           569 non-null    float64
 18  concavity_se             569 non-null    float64
 19  concave points_se        569 non-null    float64
 20  symmetry_se              569 non-null    float64
 21  fractal_dimension_se     569 non-null    float64
 22  radius_worst             569 non-null    float64
 23  texture_worst            569 non-null    float64
 24  perimeter_worst          569 non-null    float64
 25  area_worst               569 non-null    float64
 26  smoothness_worst         569 non-null    float64
 27  compactness_worst        569 non-null    float64
 28  concavity_worst          569 non-null    float64
 29  concave points_worst     569 non-null    float64
 30  symmetry_worst           569 non-null    float64
 31  fractal_dimension_worst  569 non-null    float64
 32  Unnamed: 32              0 non-null      float64
dtypes: float64(31), int64(1), object(1)
memory usage: 146.8+ KB
```

Kaggle Dataset - Breast Cancer Wisconsin (Diagnostic).

Data has **569 observations and 33 columns**.

First field is the unique 'id' number assigned to each patient.

Second field, 'diagnosis', is an indicator of the actual diagnosis ('M' = Malignant; 'B' = Benign).

There are 30 other numeric features available for prediction.

*https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

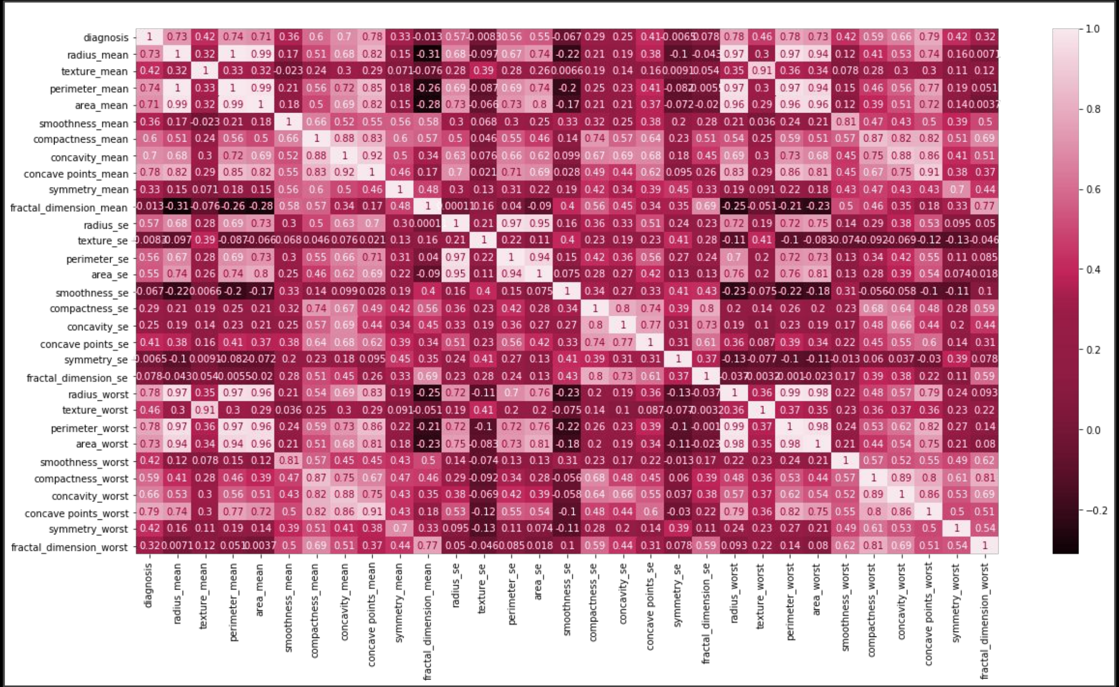
Data Processing

Deleted ‘ID’ and ‘Unnamed : 32’

Convert **‘diagnosis’** from object to int.

Find **correlation** between diagnosis and the remaining 29 fields.

Threshold of 0.75 to consider factors that are significant in deciding the diagnosis using Heatmaps

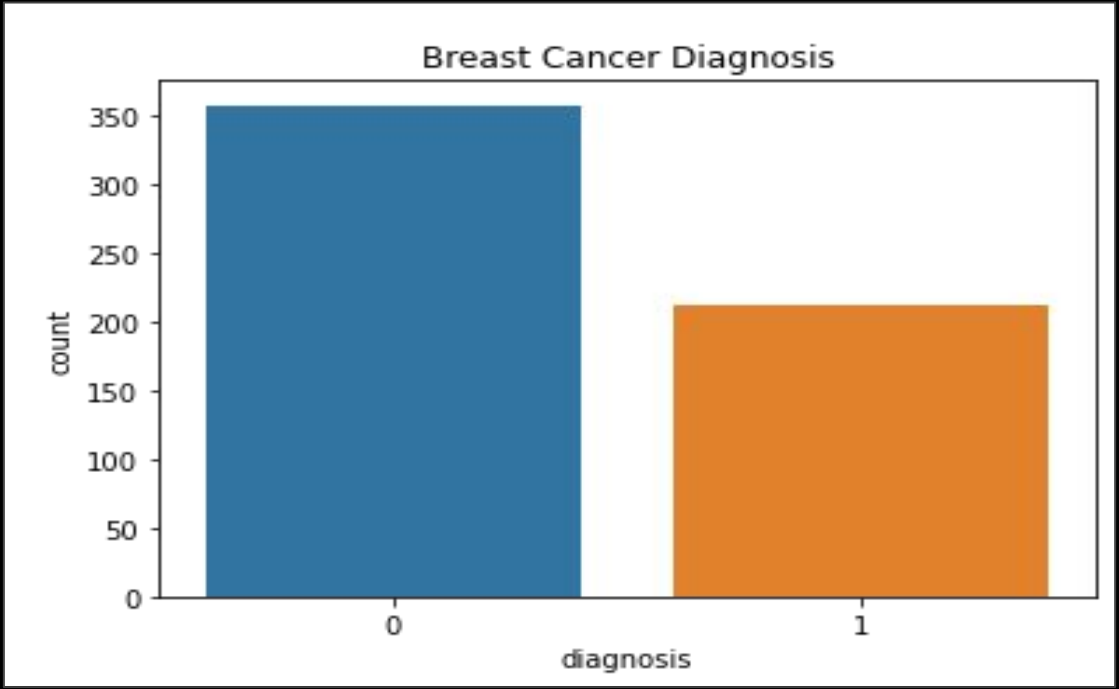


Data Visualization

Distribution of Benign and Malignant

Data is found to be:

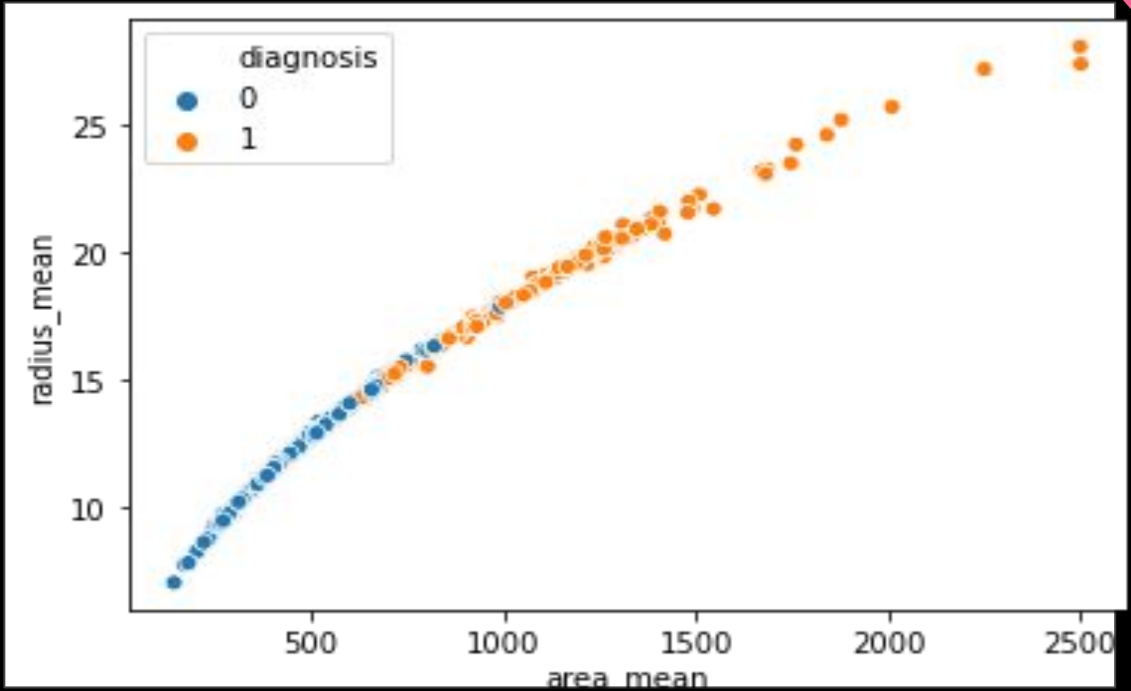
```
0    357 Refers to Benign

1    212 Refers to Malignant
```



Data Visualization

Correlation between Area mean and Radius mean:

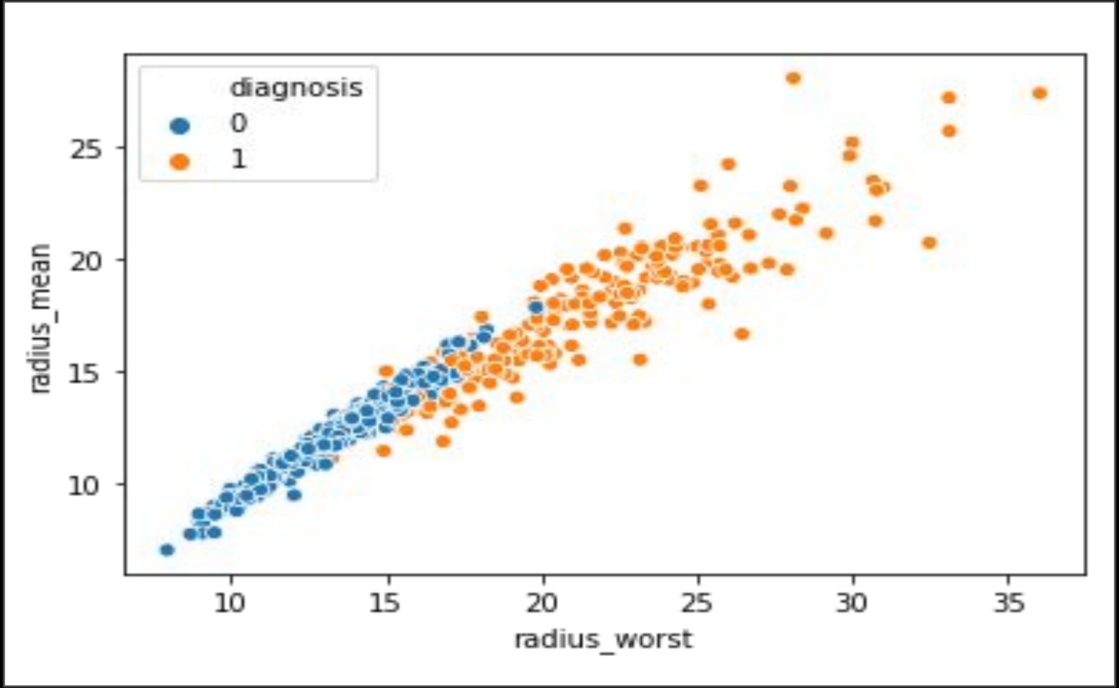
We can say that as the **area_mean** and **radius_mean** values increase there is a higher chance a female being diagnosed with Cancer.



Data Visualization

Correlation between radius worst and Radius mean:

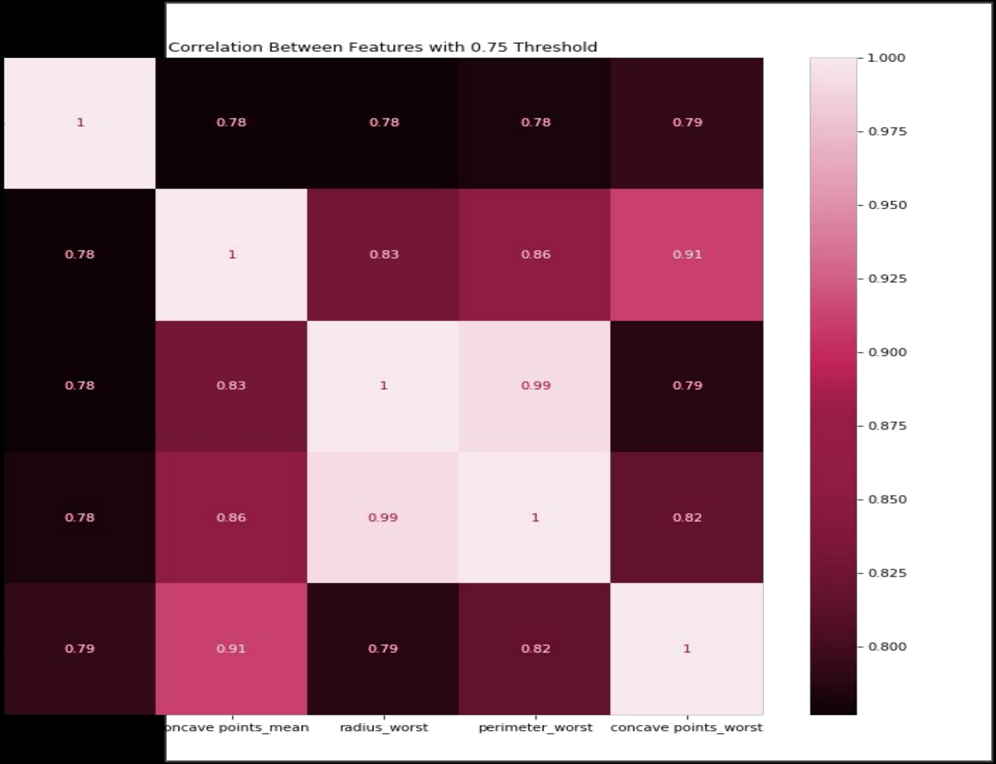
we can say that as the **radius_worst** and **radius_mean** values increase there is a higher chance a female being diagnosed with **Cancer**.



Dimensionality Reduction

Highly correlated features with respect to radius_ mean are dropped

Radius_mean is highly correlated with p**erimeter_mean, area_mean, radius_worst, perimeter_worst and area_worst. So we can drop them and use only radius_mean.**



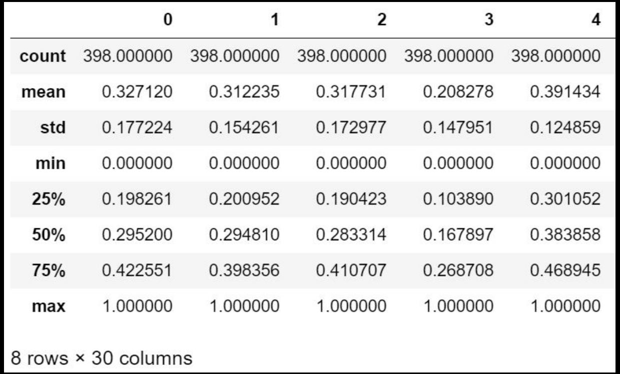
Data Processing: Feature Scaling

Most ML algos use Euclidean distance between two points.

Necessary to bring all features to the same level of magnitude.

- Normalize
- Standardize

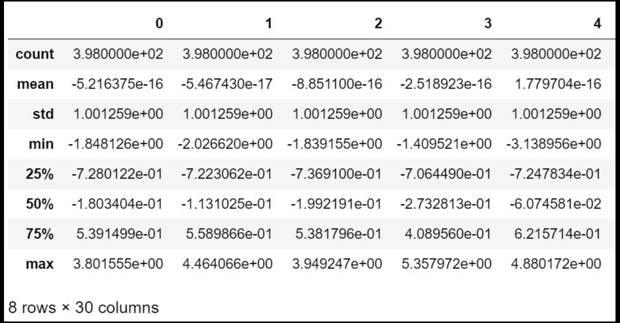
Normalized



	0	1	2	3	4
count	398.000000	398.000000	398.000000	398.000000	398.000000
mean	0.327120	0.312235	0.317731	0.208278	0.391434
std	0.177224	0.154261	0.172977	0.147951	0.124859
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.198261	0.200952	0.190423	0.103890	0.301052
50%	0.295200	0.294810	0.283314	0.167897	0.383858
75%	0.422551	0.398356	0.410707	0.268708	0.468945
max	1.000000	1.000000	1.000000	1.000000	1.000000

8 rows × 30 columns

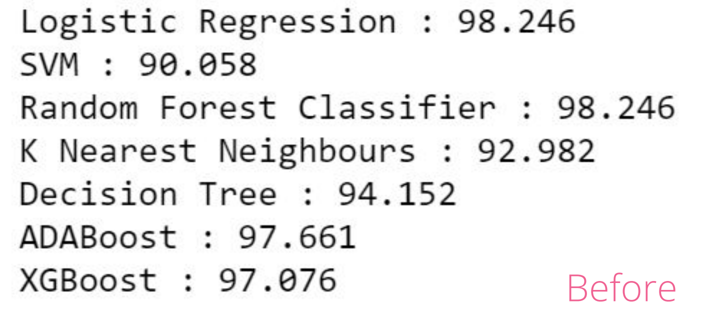
Standardized



	0	1	2	3	4
count	3.980000e+02	3.980000e+02	3.980000e+02	3.980000e+02	3.980000e+02
mean	-5.216375e-16	-5.467430e-17	-8.851100e-16	-2.518923e-16	1.779704e-16
std	1.001259e+00	1.001259e+00	1.001259e+00	1.001259e+00	1.001259e+00
min	-1.848126e+00	-2.026620e+00	-1.839155e+00	-1.409521e+00	-3.138956e+00
25%	-7.280122e-01	-7.223062e-01	-7.369100e-01	-7.064490e-01	-7.247834e-01
50%	-1.803404e-01	-1.131025e-01	-1.992191e-01	-2.732813e-01	-6.074581e-02
75%	5.391499e-01	5.589866e-01	5.381796e-01	4.089560e-01	6.215714e-01
max	3.801555e+00	4.464066e+00	3.949247e+00	5.357972e+00	4.880172e+00

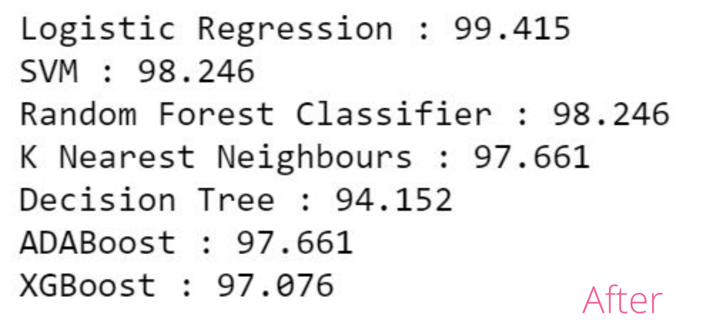
8 rows × 30 columns

Modeling



```
Logistic Regression : 98.246
SVM : 90.058
Random Forest Classifier : 98.246
K Nearest Neighbours : 92.982
Decision Tree : 94.152
ADABoost : 97.661
XGBoost : 97.076
```

Before

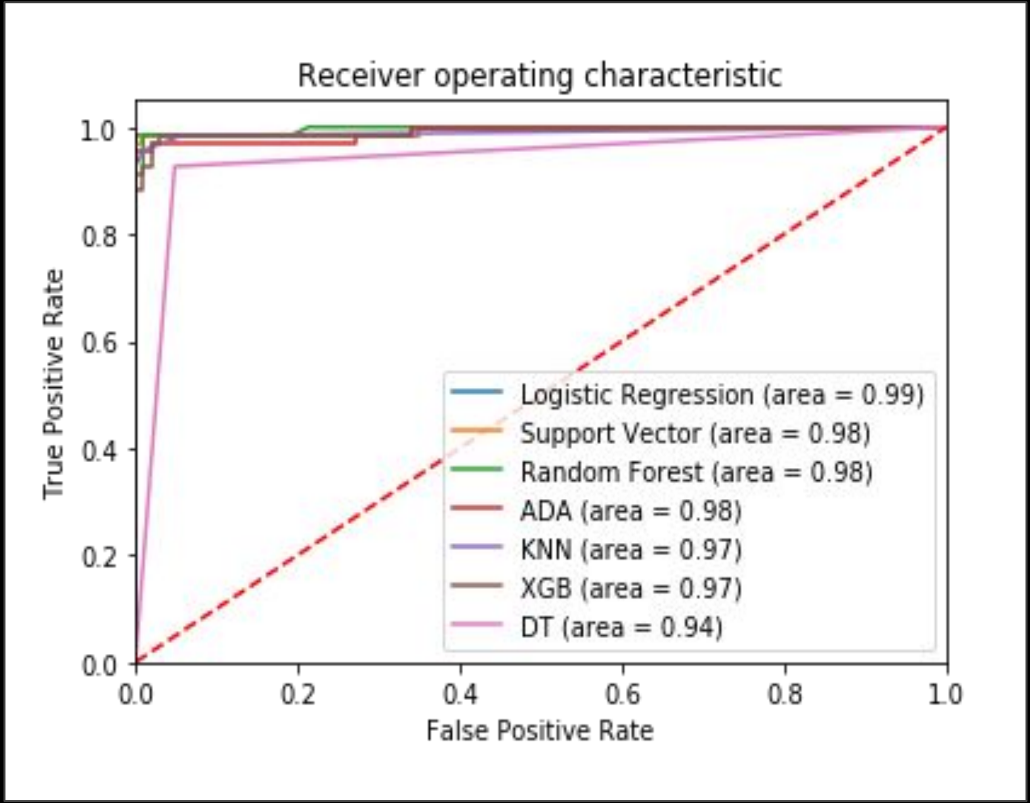


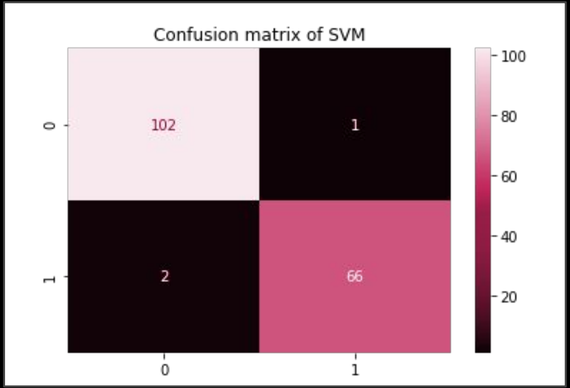
```
Logistic Regression : 99.415
SVM : 98.246
Random Forest Classifier : 98.246
K Nearest Neighbours : 97.661
Decision Tree : 94.152
ADABoost : 97.661
XGBoost : 97.076
```

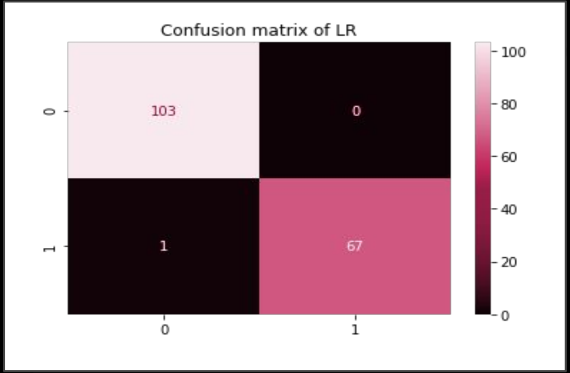
After

1. Decision Tree and Random Forest Classifier are insensitive to feature scaling.
2. Linear Regression, KNN and SVM are sensitive to feature scaling.
3. SVM and Logistic Regression models gives us the highest accuracy.

Analysis


Receiver operating characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (area = 0.99)
Support Vector (area = 0.98)
Random Forest (area = 0.98)
ADA (area = 0.98)
KNN (area = 0.97)
XGB (area = 0.97)
DT (area = 0.94)


Confusion matrix of SVM
102
1
2
66
0
1
0
1
100
80
60
40
20


Confusion matrix of LR
103
0
1
67
0
1
0
1
100
80
60
40
20
0

Conclusion

Like any other cancer, early detection of breast cancer is paramount in the effectiveness of the treatment.

Our models have proven to be successful, displaying an average accuracy of over 90% and the best model (Logistic Regression) has an accuracy of 99.415% considering only 4 of the 28 factors available.

In future scope of work, we could leverage big data technologies to predict breast cancer on a larger dataset and consider more factors for an even better accuracy and precision.

Use of pipeline utilities to find new solution that involve two or more methods working together in a complementary way to further reduce the false negative to zero.

THANK YOU

#breastcancerawareness