

IIT-M Certified Advanced Programmer with Data Science Mastery Program

By

GUVI - an IIT-Madras incubated company

IITMDSA DW42DW43

Final Project report on

**Predicting Breast Cancer in a patient using
SVM and Ensemble Techniques**

Submitted by,

Gururaj H C

DW42DW43

gururaj008@gmail.com +8050461594

<https://www.linkedin.com/in/gururaj-hc-machine-learning-enthusiast/>

Title: Predicting Breast Cancer in a patient

Abstract:

Breast cancer represents one of the diseases that make a high number of deaths every Year. It is the most common type of all cancers and the main cause of women's deaths Worldwide. Classification and data mining methods are an effective way to classify data. Especially in the medical field, where those methods are widely used in diagnosis and analysis to make decisions.

Problem Statement:

Given the details of cell nuclei taken from breast mass, predict whether or not a patient has breast cancer using the Ensembling Techniques. Perform necessary exploratory data analysis before building the model and evaluate the model based on performance metrics other than model accuracy.

About the dataset:

The dataset is comprising of 569 instances with a total of 31 features along with a target (diagnosis) which is either B or M. B stands for 'Benign' meaning that the cells are not harmful or there is no cancer and M stands for 'Malignant' meaning that the patient has cancer and the cells have a harmful effect.

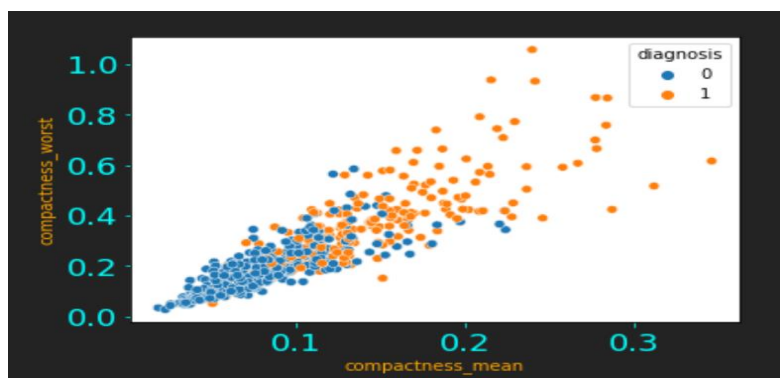
Data Pre-preprocessing:

Out of 569 instances the targets bearing 'B' are 357 and 'M' are 212. Hence, dataset is not imbalanced. All the features are of float datatype.

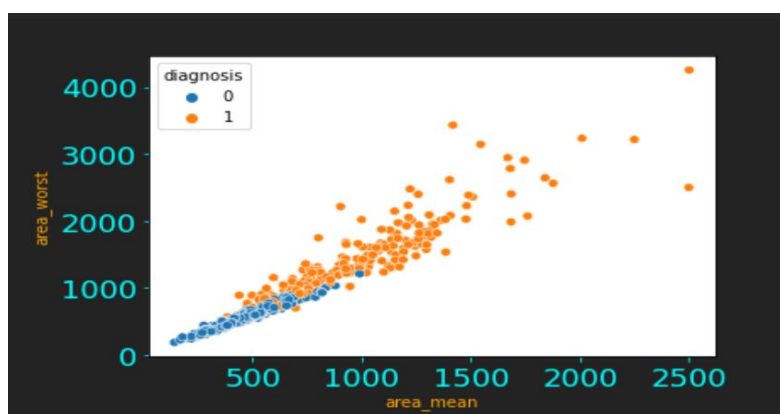
With the histplot of the dataset it is evident that only a few of the features are normally distributed while, most of them are right skewed. Hence, calling for feature scaling. The boxplot indicates many features having outliers, so choosing **Standardization** over Normalization.

Exploratory data analysis

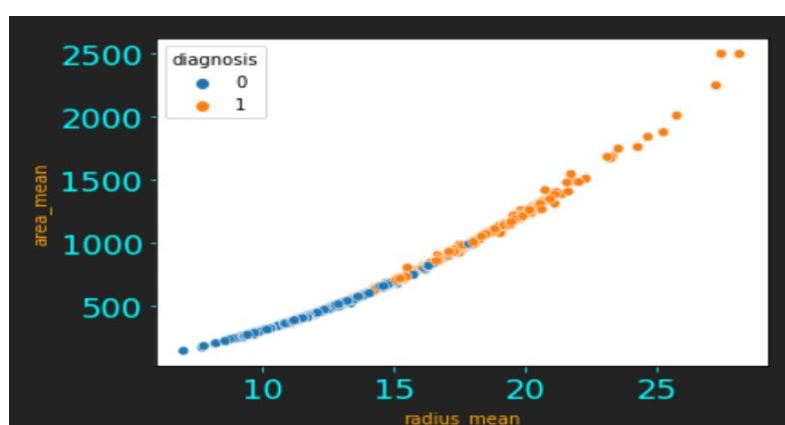
Relationship between the important features as a pair to explore their impact on target



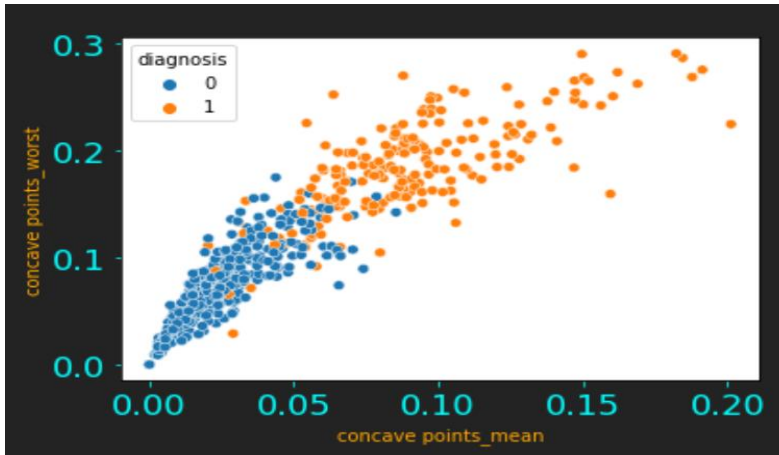
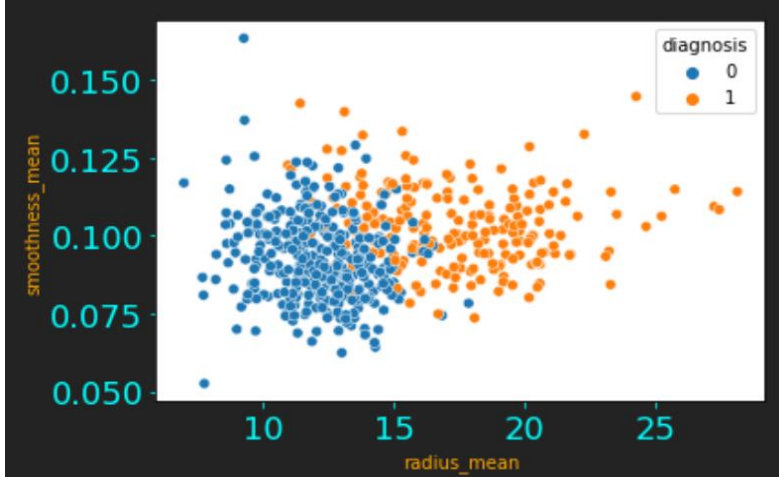
As evident from the plot, with the increase in compactness_mean beyond 0.11 and compactness_worst beyond 0.3, the chances of being Malignant(1) increases significantly



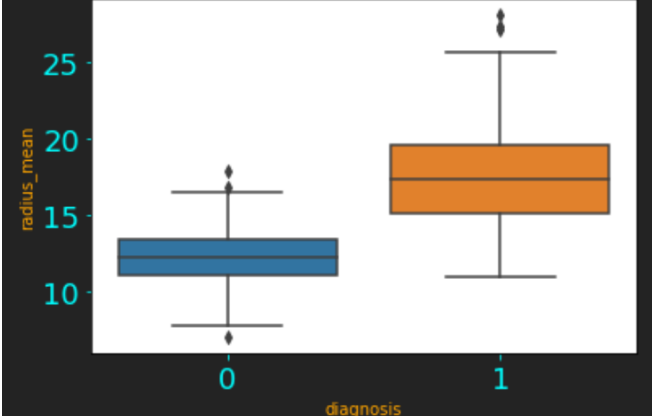
With the increase in area_mean beyond 1000 and area_worst beyond 800, the chances of being Malignant(1) increases

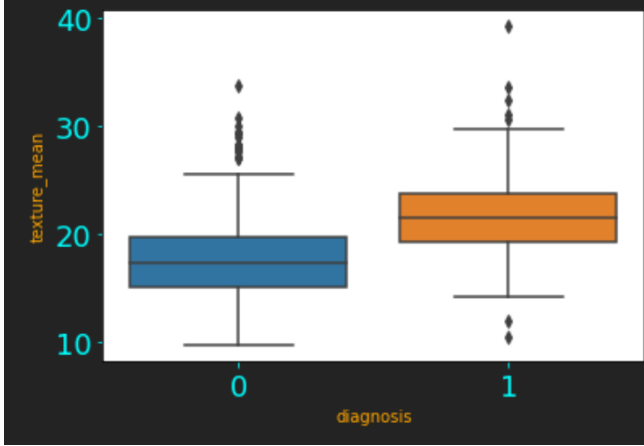
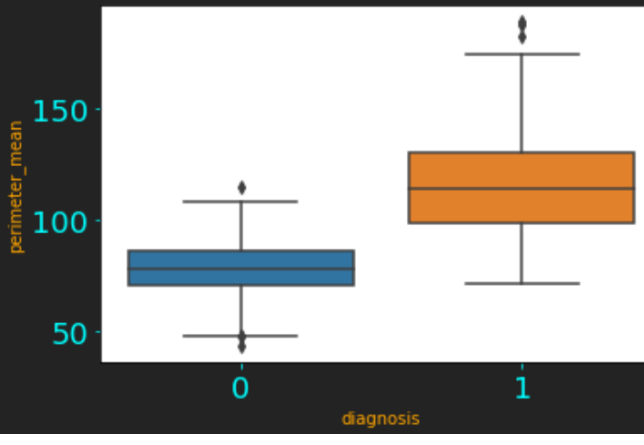
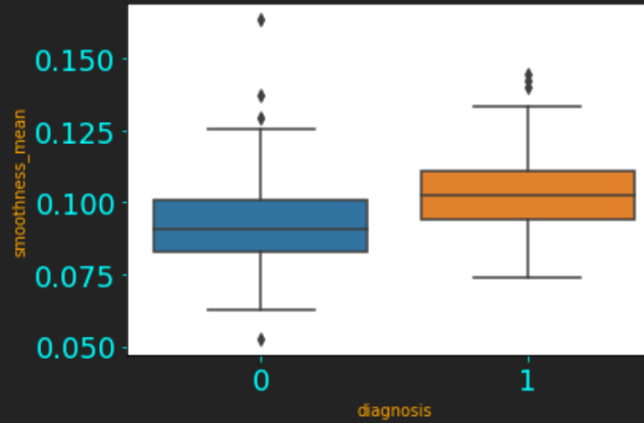


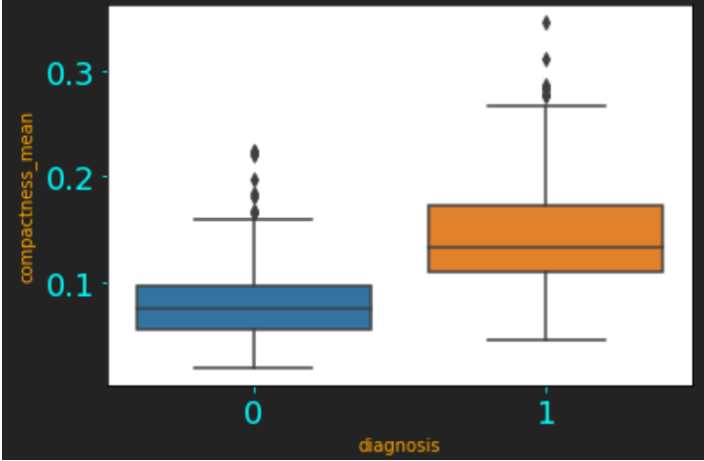
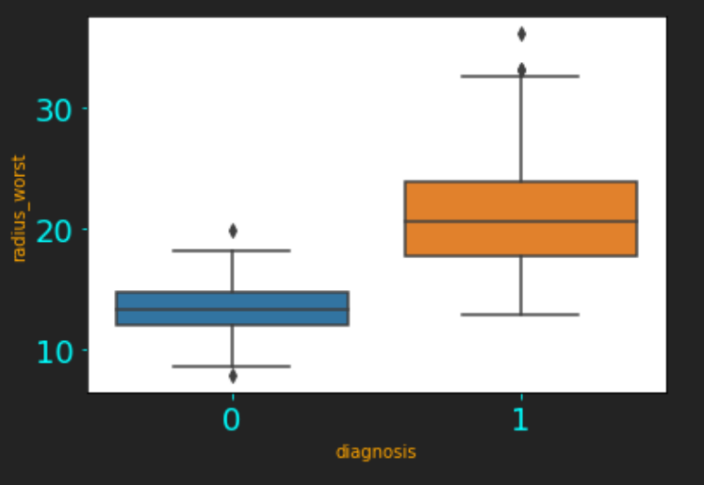
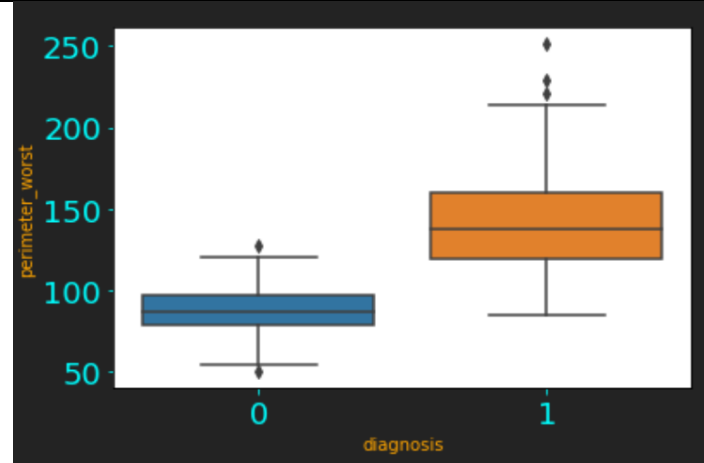
With the increase in radius_mean beyond 17 and area_mean beyond 800, the chances of being Malignant(1) increases

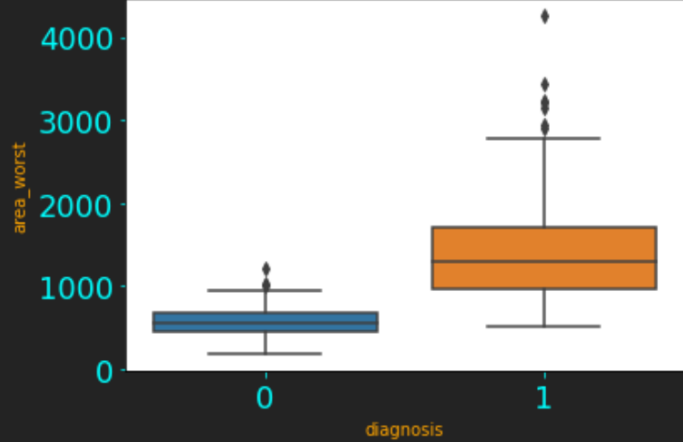
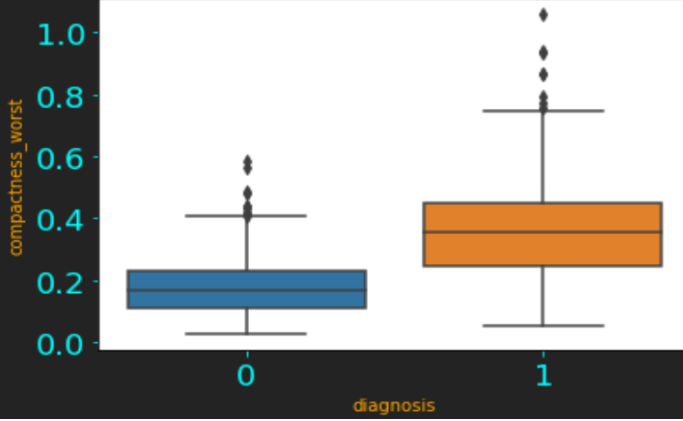
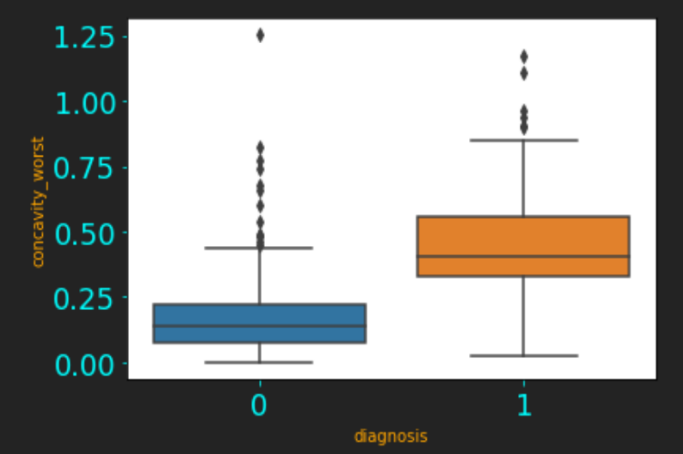
	<p>With the increase in concave_points_mean beyond 0.07 and concave_points_worst beyond 0.16 the chances of being Malignant(1) increases</p>
	<p>Regardless of the value of smoothness_mean, if the value of radius_mean is beyond 14, the chances of being Malignant(1) increases</p>

Relationship between the important features and target

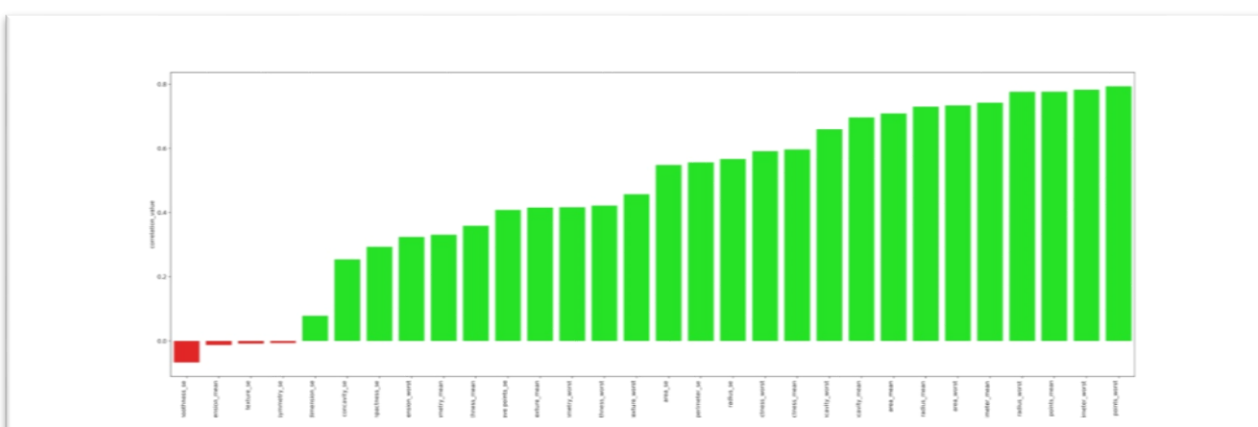
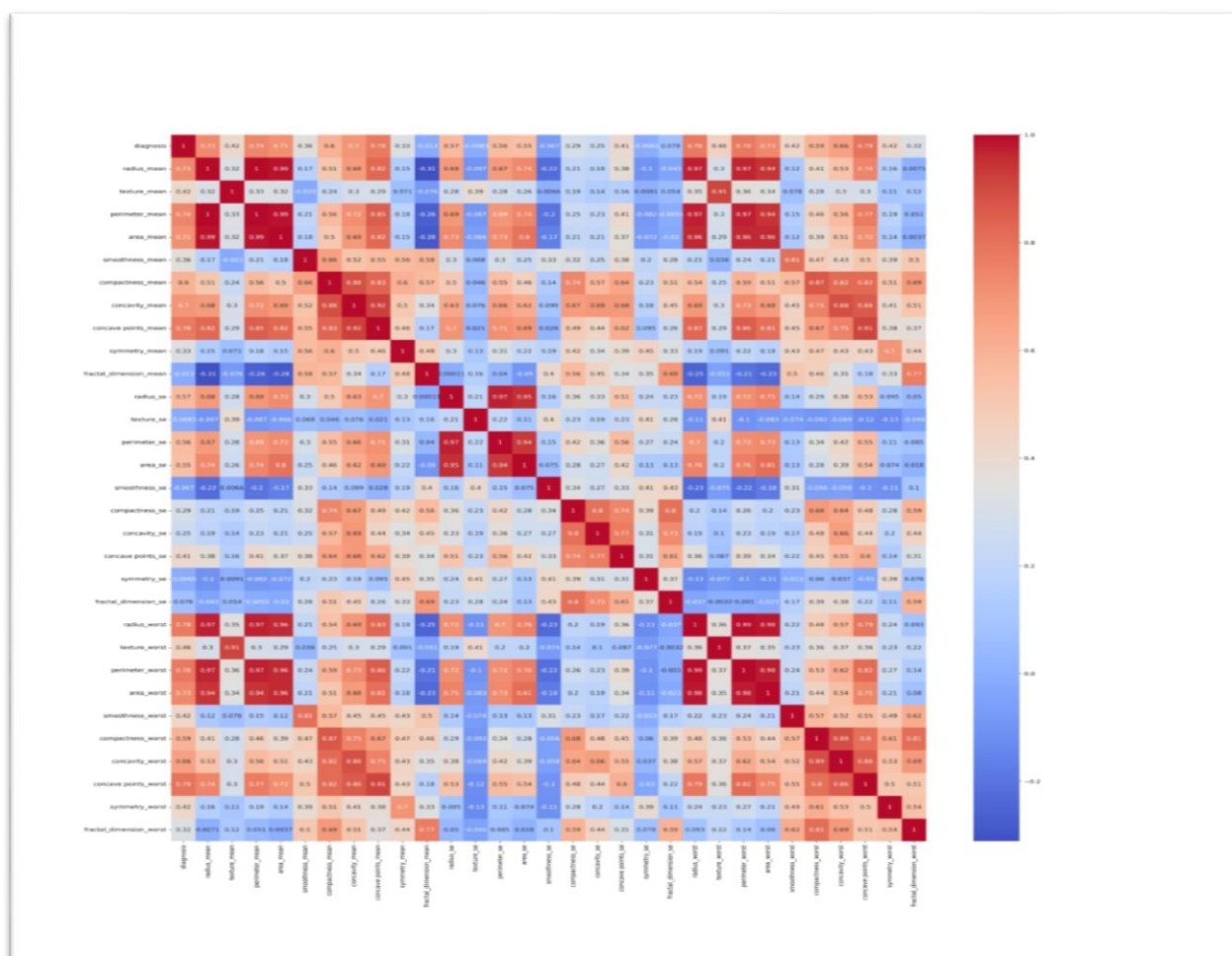
	<p>If the value of radius_mean is beyond 17 chances of being Malignant(1) increases</p>
---	---

 <p>A box plot comparing the texture_mean for two diagnosis categories: 0 (benign) and 1 (malignant). The y-axis is labeled 'texture_mean' and ranges from 10 to 40. The x-axis is labeled 'diagnosis' with categories 0 and 1. The box for diagnosis 0 is blue, with a median around 17. The box for diagnosis 1 is orange, with a median around 22. Whiskers extend from the boxes, and individual data points are plotted as small diamonds above and below the whiskers.</p>	<p>If the value of texture_mean is beyond 21 chances of being Malignant(1) increases</p>
 <p>A box plot comparing the perimeter_mean for two diagnosis categories: 0 (benign) and 1 (malignant). The y-axis is labeled 'perimeter_mean' and ranges from 50 to 150. The x-axis is labeled 'diagnosis' with categories 0 and 1. The box for diagnosis 0 is blue, with a median around 80. The box for diagnosis 1 is orange, with a median around 115. Whiskers extend from the boxes, and individual data points are plotted as small diamonds above and below the whiskers.</p>	<p>If the value of texture_mean is beyond 21 chances of being Malignant(1) increases</p>
 <p>A box plot comparing the smoothness_mean for two diagnosis categories: 0 (benign) and 1 (malignant). The y-axis is labeled 'smoothness_mean' and ranges from 0.050 to 0.150. The x-axis is labeled 'diagnosis' with categories 0 and 1. The box for diagnosis 0 is blue, with a median around 0.09. The box for diagnosis 1 is orange, with a median around 0.105. Whiskers extend from the boxes, and individual data points are plotted as small diamonds above and below the whiskers.</p>	<p>If the value of smoothness_mean is beyond 0.1 chances of being Malignant(1) increases</p>

 <p>A box plot showing the distribution of 'compactness_mean' for two diagnosis categories: 0 (benign) and 1 (malignant). The y-axis ranges from 0.1 to 0.3. The box for diagnosis 0 is blue, with a median around 0.08. The box for diagnosis 1 is orange, with a median around 0.14. Whiskers extend from the boxes, and individual data points are plotted as black dots above the boxes.</p>	<p>If the value of compactness_mean is beyond 0.12 chances of being Malignant(1) increases</p>
 <p>A box plot showing the distribution of 'radius_worst' for two diagnosis categories: 0 (benign) and 1 (malignant). The y-axis ranges from 10 to 30. The box for diagnosis 0 is blue, with a median around 13. The box for diagnosis 1 is orange, with a median around 21. Whiskers extend from the boxes, and individual data points are plotted as black dots above the boxes.</p>	<p>If the value of radius_worst is beyond 18 chances of being Malignant(1) increases</p>
 <p>A box plot showing the distribution of 'perimeter_worst' for two diagnosis categories: 0 (benign) and 1 (malignant). The y-axis ranges from 50 to 250. The box for diagnosis 0 is blue, with a median around 90. The box for diagnosis 1 is orange, with a median around 140. Whiskers extend from the boxes, and individual data points are plotted as black dots above the boxes.</p>	<p>If the value of perimeter_worst is beyond 130 chances of being Malignant(1) increases</p>

 <p>A box plot comparing the 'area_worst' variable for two diagnosis categories: 0 (benign) and 1 (malignant). The y-axis ranges from 0 to 4000. The box for diagnosis 0 is blue, with a median around 600. The box for diagnosis 1 is orange, with a median around 1300. Whiskers extend from the boxes, and individual data points are plotted as black dots above the whiskers.</p>	<p>If the value of area_worst is beyond 1000 chances of being Malignant(1) increases</p>
 <p>A box plot comparing the 'compactness_worst' variable for two diagnosis categories: 0 (benign) and 1 (malignant). The y-axis ranges from 0.0 to 1.0. The box for diagnosis 0 is blue, with a median around 0.2. The box for diagnosis 1 is orange, with a median around 0.35. Whiskers extend from the boxes, and individual data points are plotted as black dots above the whiskers.</p>	<p>If the value of compactness_worst is beyond 0.3 chances of being Malignant(1) increases</p>
 <p>A box plot comparing the 'concavity_worst' variable for two diagnosis categories: 0 (benign) and 1 (malignant). The y-axis ranges from 0.00 to 1.25. The box for diagnosis 0 is blue, with a median around 0.15. The box for diagnosis 1 is orange, with a median around 0.4. Whiskers extend from the boxes, and individual data points are plotted as black dots above the whiskers.</p>	<p>If the value of concavity_worst is beyond 0.3 chances of being Malignant(1) increases</p>

Heatmap showing the relationship between features and target using `df.corr ()` function



As seen from the plot above 4 features are negatively correlated with the target, while the rest are positively correlated. Nearly 10 features show a strong positive correlation with the final outcome.

Features with the negative correlation doesn't have much of an impact on the target

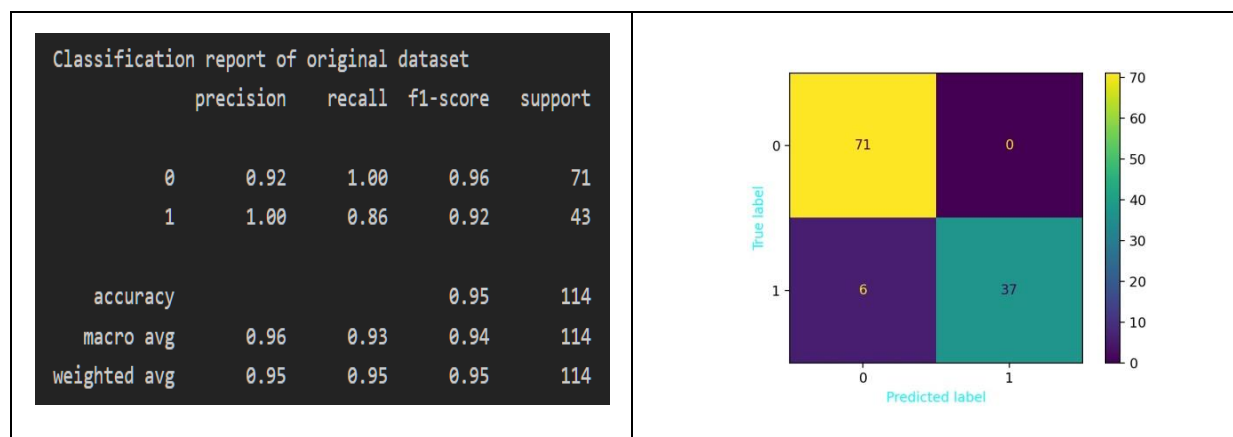
Features having negative correlation with the target		
	parameter	correlation_value
0	smoothness_se	-0.067016
1	fractal_dimension_mean	-0.012838
2	texture_se	-0.008303
3	symmetry_se	-0.006522

Features with positive correlation do have a large impact on the target as evident below,

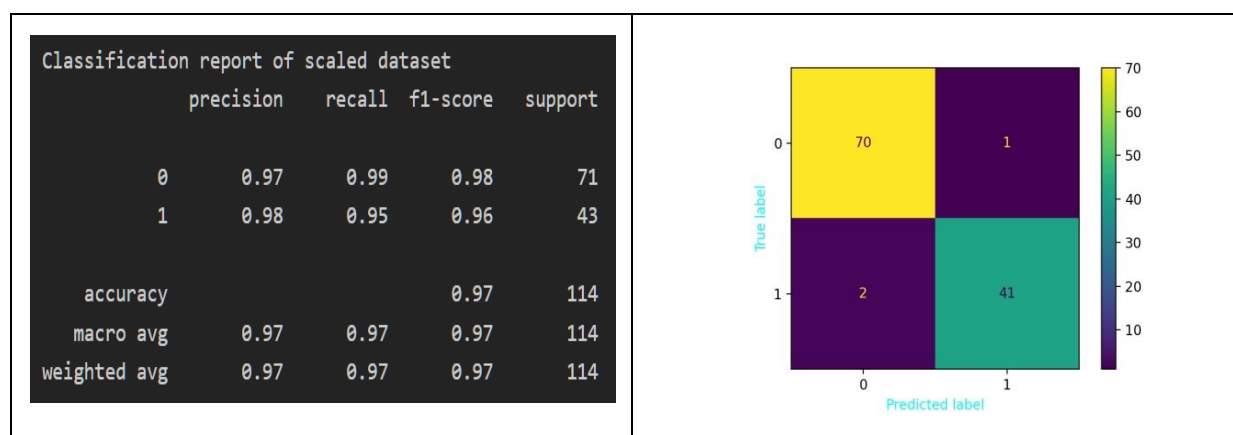
Features having positive correlation with the target		
	parameter	correlation_value
0	fractal_dimension_se	0.077972
1	concavity_se	0.253730
2	compactness_se	0.292999
3	fractal_dimension_worst	0.323872
4	symmetry_mean	0.330499
5	smoothness_mean	0.358560
6	concave points_se	0.408042
7	texture_mean	0.415185
8	symmetry_worst	0.416294
9	smoothness_worst	0.421465
10	texture_worst	0.456903
11	area_se	0.548236
12	perimeter_se	0.556141
13	radius_se	0.567134
14	compactness_worst	0.590998
15	compactness_mean	0.596534
16	concavity_worst	0.659610
17	concavity_mean	0.696360
18	area_mean	0.708984
19	radius_mean	0.730029
20	area_worst	0.733825
21	perimeter_mean	0.742636
22	radius_worst	0.776454
23	concave points_mean	0.776614
24	perimeter_worst	0.782914
25	concave points_worst	0.793566

Before proceeding with feature scaling and hyper parameter tuning, let us get the F1 score of the dataset to set a benchmark first. As seen below the original dataset has an F1 score of 0.96 for not having cancer and 0.92 for having cancer with an overall accuracy of 0.95.

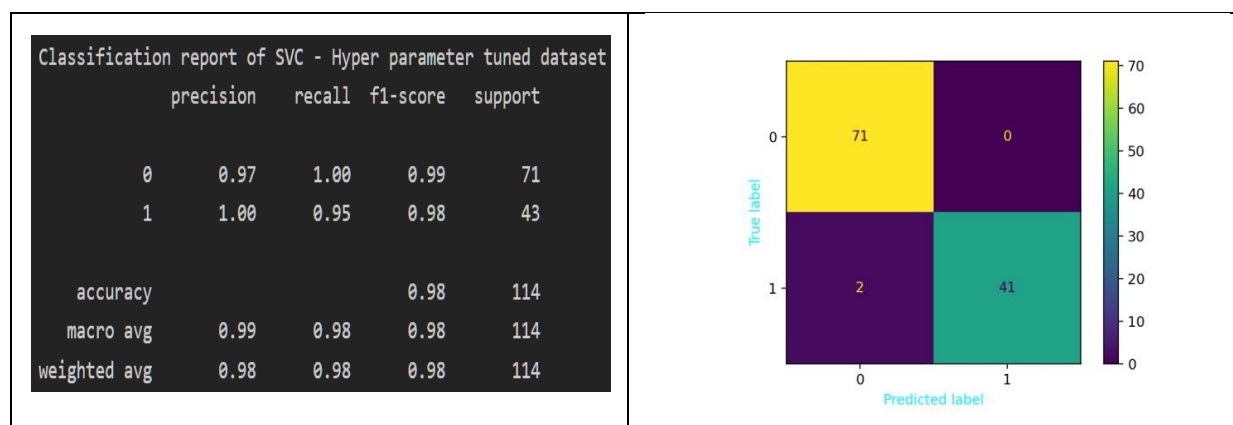
Report on GUVI final project: Predicting Breast Cancer in a patient



After feature scaling both the F1 score and accuracy have improved as noticed below,



Now using Support Vector Classifier with tuned hyper-parameters,

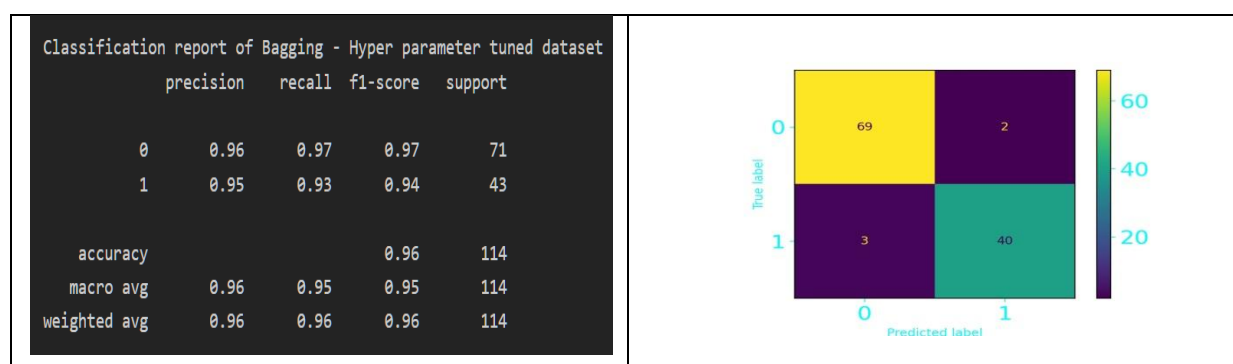


Report on GUVI final project: Predicting Breast Cancer in a patient

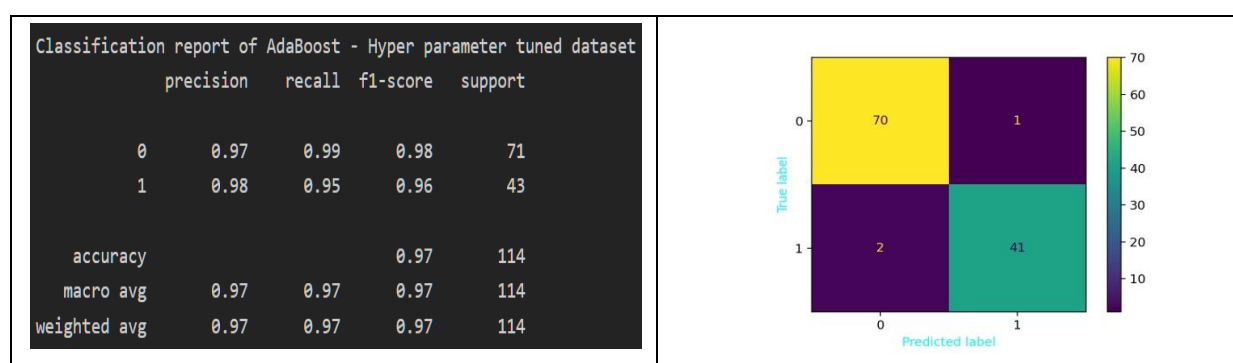
Now using Random Forest Classifier with tuned hyper-parameters,



Now using Bagging Classifier with tuned hyper-parameters,

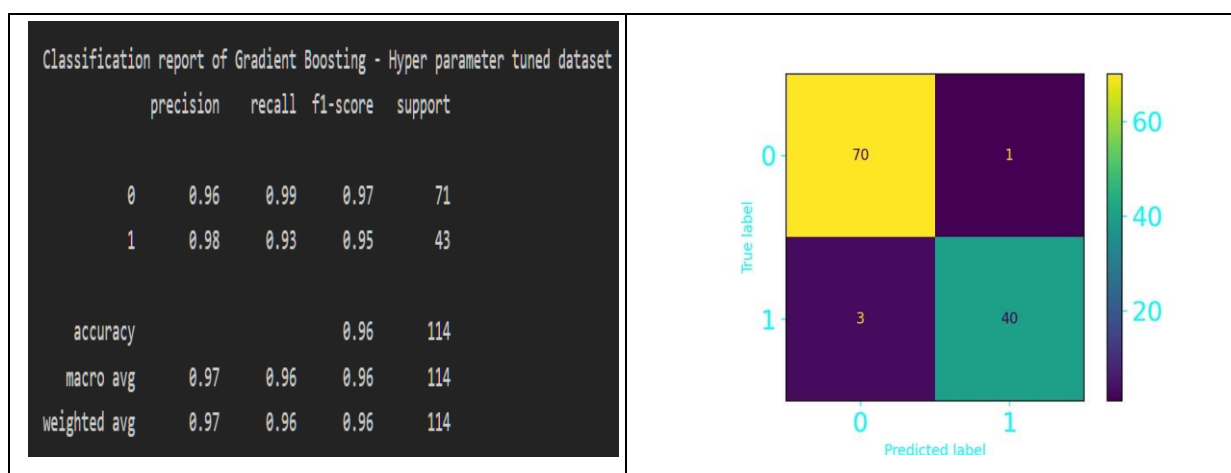


Now using Ada Boost Classifier with tuned hyper-parameters,



Report on GUVI final project: Predicting Breast Cancer in a patient

Now using Gradient Boost Classifier with tuned hyper-parameters,



Now using Extreme Gradient Boost Classifier with tuned hyper-parameters,

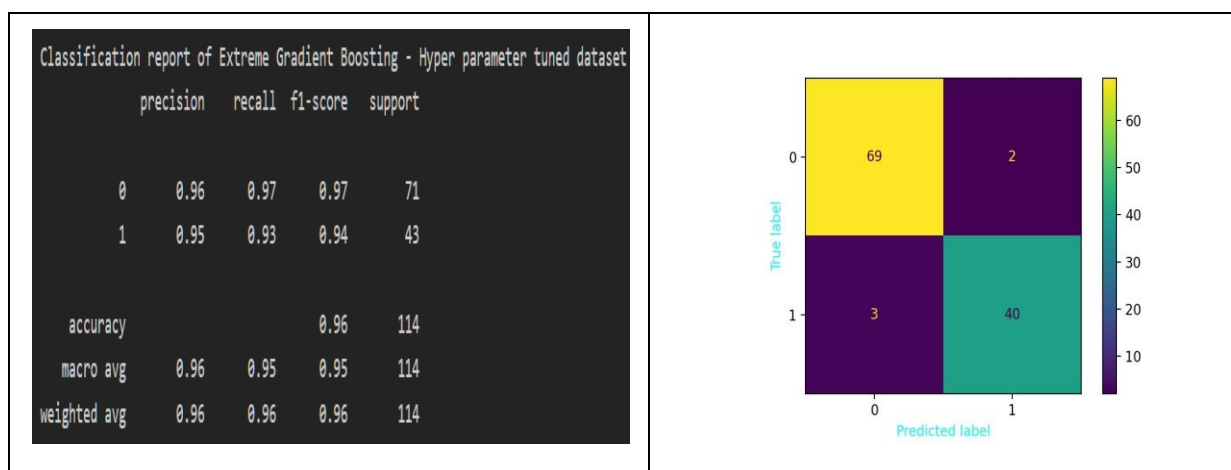


Chart showing the performance of various models on the dataset

		False_Positive	False_Negative	Accuracy	Precision	Recall	Specificity	F1_Score
Index								
Original dataset	0	6		0.95	1.00	0.86	1.00	0.92
Scaled dataset	1	2		0.97	0.98	0.95	0.99	0.96
SVC - Tuned dataset	0	2		0.98	1.00	0.95	1.00	0.97
RFC - Tuned dataset	1	3		0.96	0.98	0.93	0.99	0.95
Bagging Classifier - Tuned dataset	2	3		0.96	0.95	0.93	0.97	0.94
Ada Boost - Tuned dataset	1	2		0.97	0.98	0.95	0.99	0.96
Gradient Boost - Tuned dataset	1	3		0.96	0.98	0.93	0.99	0.95
Extreme Gradient Boost - Tuned dataset	2	3		0.96	0.95	0.93	0.97	0.94

Significance of False Positive:

Since it is a medical dataset, False positive means wrongly predicting the individual to be having a cancer but in reality the individual doesn't. Usually people go for a second opinion. Hence, it is not much of life threatening significance. That being said SVC is the only model which didn't produce any False Positives.

Significance of False Negative:

Since it is a medical dataset, False Negative means wrongly predicting the individual to be not having a cancer but in reality the individual does have cancer, which is a very dangerous and life threatening scenario. SVC and Ada Boost models performed well compared to other models, predict only 02 False Negatives out of 114 predictions accounting for about 1.75% wrong prediction.

Significance of Accuracy:

Accuracy is the ratio of the total number of correct predictions and the total number of predictions. SVC is the clear winner here with accuracy of 98%.

Significance of Precision:

It is the measure of patients correctly identify as having cancer out of all the patients actually having it. SVC is the only model which predicted 100% efficiently.

Significance of Recall:

For all the patients who actually have cancer, recall tells us how many we correctly identified as having a cancer. SVC and Ada Boost models lead the way with 95% efficiency.

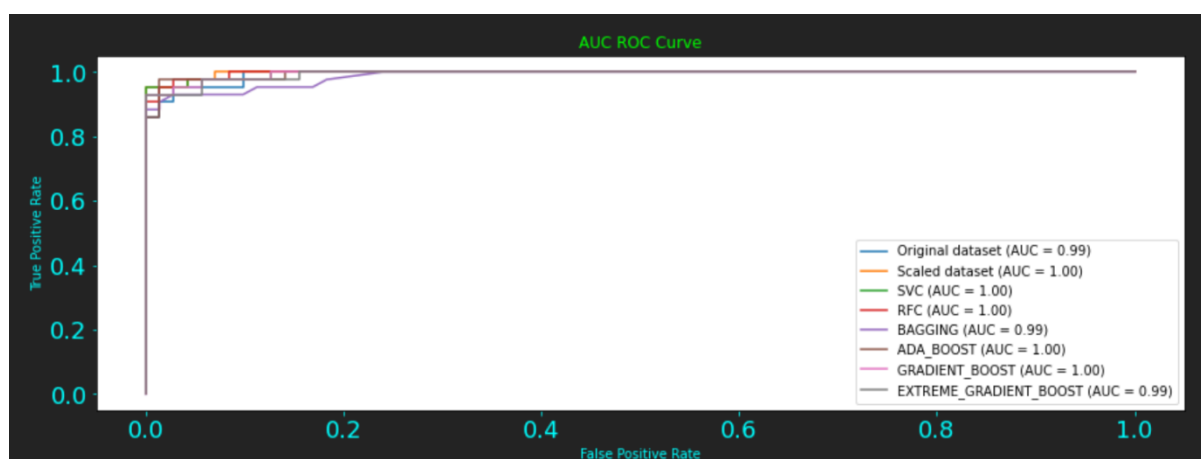
Significance of Specificity:

it is the measure of the number of cases where the model correctly predicts that the patient does not have cancer from all the patients who actually didn't have cancer. SVC performed better than other models with 100% efficiency.

Significance of F1_Score:

F1-score is the Harmonic mean of the Precision and Recall. This is easier to work with since now, instead of balancing precision and recall, we can just aim for a good F1-score, which would also indicate good Precision and a good Recall value. Again SVC is the clear winner here with a F1 Score of 0.97.

ROC Curve



Receiver Operating Characteristic Curve (ROC Curve) It is the plot between the TPR(y-axis) and FPR(x-axis). The more the area under the curve the better. SVC, RFC, ADA_BOOST, GRADIENT_BOOST have maximum area under the curve.

The verdict: **SVC outperformed Ensemble models like (Random Forest, Bagging, AdaBoost, GradientBoost and even ExtremeGradientBoost Classifier algorithms) with a superior F1_Score, 0 False Positives and only 2 False Negatives for the entire dataset.**