Breast Cancer Prediction Milestone: Model Performance Evaluation and Interpretation

Group 25

Student 1: Siva Vasantha Harika Mangu
Student 2: Raaga Sindhu Mangalagiri

mangu.s@northeastern.edu
mangalagiri.r@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Siva Vasantha Harika Mangu

Signature of Student 2: Raaga Sindhu Mangalagiri

Submission Date: 03/24/2023

Project Proposal

IE 7275: Data Mining in Engineering

Problem Statement:

About 1 in 8 women will develop invasive breast cancer over the course of their lifetime. And every year around 40,000 women in United States alone, are dying from breast cancer. It starts when cells in breast begin to grow out of control. These cells usually form tumors that can be seen via X-ray often felt as lumps in the breast area. And one of the shocking revelations can be said that it is found in women who don’t show any symptoms. It is mostly occurring disease among women and in specific elderly age groups. They cannot prevent or control their risk of cancer, and it cannot even be recognized at early stage due to lack of symptoms and abnormalities they face.

Hence, by building a machine learning model, we can be able to predict the risks of breast cancer among women at early age, based on different attributes acting as a contributing factor to the disease.

Problem Definition:

Given a set of patient data including demographic information, medical history, and medical imaging results, the goal is to develop a model or system that can accurately predict the likelihood of a patient developing breast cancer. The challenge is to classify these tumors into malignant(cancerous) or benign(non-cancerous). The model should be able to handle missing or incomplete data and be able to generalize well to new unseen cases. Additionally, the model should be interpretable and provide insights on the most important features/variables that contribute to the prediction. The goal is to improve the early detection and prevention of breast cancer by identifying high-risk individuals and providing them with the necessary interventions, resulting in better outcomes for patients.

Data Sources:

Kaggle

[Breast Cancer Dataset | Kaggle]

Data Description:

Breast cancer data set consists of 32 attributes and 570 records in total, in which 30 attributes are contributing as predictors and the attribute diagnosis is the response variable which predicts whether the patient is Benign or malignant. The Unique ID is the primary key which is the not the predictor either or a response variable. The predictors such as radius_mean,texture_mean,perimeter_mean,area_mean,smoothness_mean,compactness_mean,concavity_mean,concavepoints_mean,symmetry_mean,fractal_dimension_mean,radius_se,texture_se,perimeter_se,area_se,smoothness_se,compactness_worst,concavity_worst,concavepoints_worst,symmetry_worst,fractal_dimension_worst.Some of the other risk factors which can be considered as predictors are age, family history, certain genetic mutations, and certain lifestyle factors such as alcohol consumption and lack of physical activity. Symptoms of breast cancer include a lump or thickening in the breast tissue, changes in the size or shape of the breast, and changes to the skin on the breast such as redness or dimpling.

Insights

List of Predictors:

The dataset we are working on contains various measurements related to breast cancer diagnosis. Here is an overview of the attributes(columns) we have in the dataset:

1. radius_mean: mean of distances from the center to points on the perimeter of the tumor.
2. texture_mean: standard deviation of gray-scale values
3. perimeter_mean: perimeter of the tumor
4. area_mean: area of the tumor
5. smoothness_mean: local variation in radius lengths
6. compactness_mean: perimeter^2/area – 1.0
7. concave points_mean: number of concave portions of the contour
8. symmetry_mean: symmetry of the tumor
9. fractal_dimension_mean: “coastline approximation” - 1
10. radius_se: standard error of the mean of distances from the center to points on the perimeter

11. texture_se: standard error of gray-scale values
12. perimeter_se: standard error of the perimeter
13. area_se: standard error of the area
14. smoothness_se: standard error of local variation in radius lengths
15. compactness_se: standard error of perimeter^2/area – 1.0
16. concavity_se: standard error of number of concave portions of the contour
17. concave points_se: standard error of number of concave portions of the contour
18. symmetry_se: standard error of symmetry of the tumor
19. fractal_dimension_se: standard error of "coastline approximation" – 1
20. radius_worst: "worst" or largest mean value from the mean of distances from the center to points on the perimeter
21. texture_worst: "worst" or largest standard deviation of gray-scale values
22. perimeter_worst: "worst" or largest perimeter of the tumor
23. area_worst: "worst" or largest area of the tumor
24. smoothness_worst: "worst" or largest local variation in radius lengths
25. compactness_worst: "worst" or largest perimeter^2/area – 1.0
26. concavity_worst: "worst" or largest number of concave portions of the contour
27. concave_points_worst: "worst" or largest number of concave portions of the contour
28. symmetry_worst: "worst" or largest symmetry of the tumor
29. fractal_dimension_worst: "worst" or largest "coastline approximation" – 1

Each attribute provides different information about the tumor, such as its size, texture, shape, and smoothness. Understanding what each attribute represents helps make decision about feature selection and model performance evaluation.

Exploratory Data Analysis

To determine which attributes are highly correlated and could be useful in predicting breast cancer, we can calculate the correlation matrix of the dataset. The correlation matrix shows the correlation coefficient between each pair of attributes and find potential attributes for the better prediction, where a values of 1 indicates a positive correlation, 0 indicates no correlation, and -1 indicates a perfect negative correlation. The values greater than 0.7 are considered as highly correlated. Next, we found the correlation of these highly correlated variables with the response variable(diagnosis) to determine the targeted variables responsible for our prediction.

Dividing the dataset for better understanding and visualization:

The first 10 attributes (radius mean, texture mean, perimeter mean, area mean, smoothness mean, compactness mean, concavity mean, concave points mean, symmetry mean, and fractal dimension) are measures of the tumor's size, shape, and texture based on a digital image of a biopsy. The next 10 attributes (radius se, texture se, perimeter se, area se, smoothness se, compactness se, concavity se, concave points se, symmetry se, and fractal dimension se) are the standard errors of these same measurements, which indicate the variability or uncertainty of the estimates. The last 10 attributes (radius worst, texture worst, area worst, smoothness worst, concavity worst, concave points worst, symmetry worst, and fractal dimension worst) are the "worst" or largest values of these measurements observed in the biopsy, which indicate the aggressiveness or malignancy of the tumor. These features can be used to build a machine learning model for breast cancer prediction diagnosis. Therefore, we are concentrating on finding the most correlated attributes on 3 different sets i.e., first 10, second 10 and the last, inorder to build a machine learning model for breast can cer prediction or diagnosis.

We are finding Correlation matrix in order to find the relation between individual attributes and hence able to scale down with attributes which have high correlation, thus helping us to understand what attributes we must focus on to predict breast cancer. We have used inbuilt corr() function to find the correlation matrix of first 10 attributes. Below is the correlation matrix We have mapped the correlation values in a heatmap using seaborn in order to recognize highly correlated attributes. And then, have segregated values which are above 0.7 as highly correlated

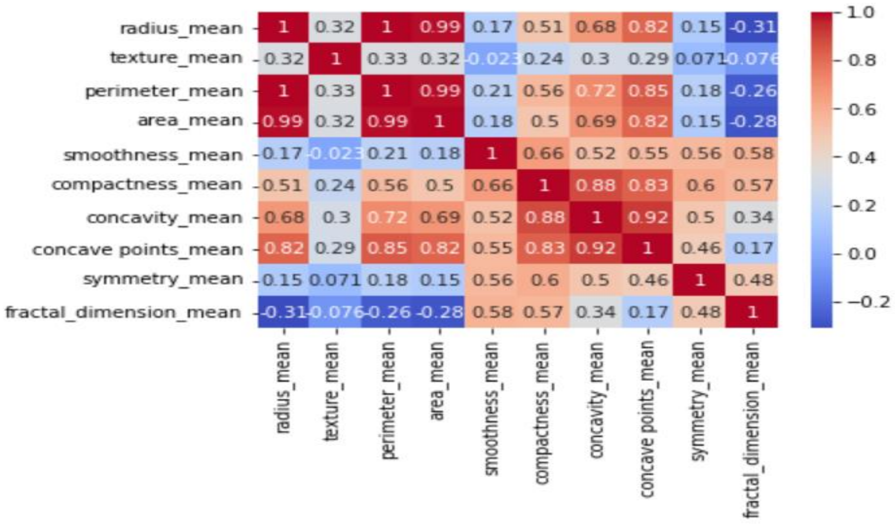


Figure1: Heatmap of correlation values of first 10 attributes

The above steps such as finding the correlation matrix and finding the attributes which are highly correlated has been repeated and plotted a heatmap and segregated the values which are above 0.7 as highly correlated values.

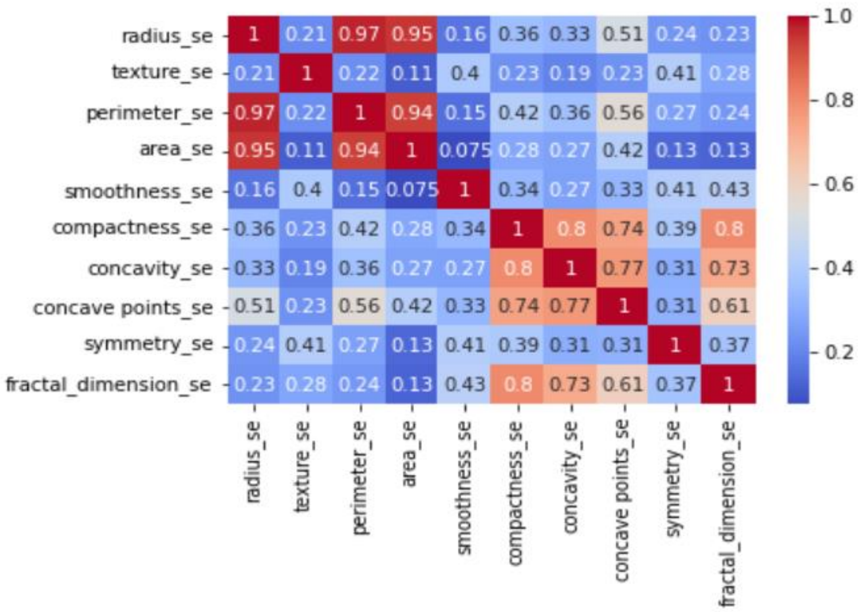
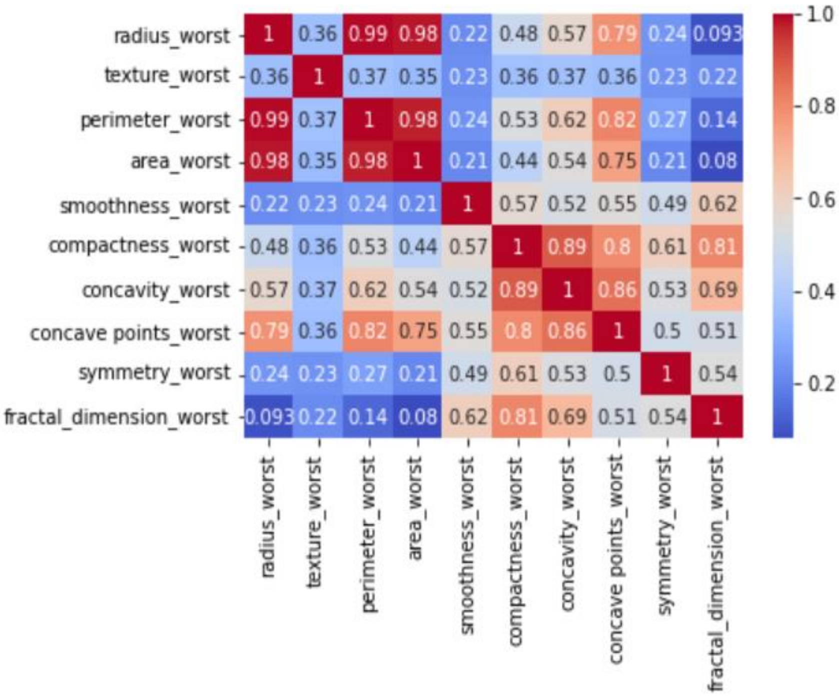


Figure 2: Heatmap of Correlation matrix of next 10 attributes

Again, the above steps such as finding the correlation matrix and finding the attributes which are highly correlated has been repeated and plotted a heatmap and segregated the values which are above 0.7 as highly correlated values.



	radius_worst	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
radius_worst	1	0.36	0.99	0.98	0.22	0.48	0.57	0.79	0.24	0.093
texture_worst	0.36	1	0.37	0.35	0.23	0.36	0.37	0.36	0.23	0.22
perimeter_worst	0.99	0.37	1	0.98	0.24	0.53	0.62	0.82	0.27	0.14
area_worst	0.98	0.35	0.98	1	0.21	0.44	0.54	0.75	0.21	0.08
smoothness_worst	0.22	0.23	0.24	0.21	1	0.57	0.52	0.55	0.49	0.62
compactness_worst	0.48	0.36	0.53	0.44	0.57	1	0.89	0.8	0.61	0.81
concavity_worst	0.57	0.37	0.62	0.54	0.52	0.89	1	0.86	0.53	0.69
concave points_worst	0.79	0.36	0.82	0.75	0.55	0.8	0.86	1	0.5	0.51
symmetry_worst	0.24	0.23	0.27	0.21	0.49	0.61	0.53	0.5	1	0.54
fractal_dimension_worst	0.093	0.22	0.14	0.08	0.62	0.81	0.69	0.51	0.54	1

Figure 3: Heatmap of Correlation values of last 10 attributes

After comparing all the values, the columns which have highest correlation values among first 10 attributes are "perimeter_mean" and "concave points_mean". Hence, we have plotted a pair plot of these columns to check the malignant and benign data points, in order to crosscheck the correlation concept.

```
<seaborn.axisgrid.PairGrid at 0x7fbb8a7cb460>
```

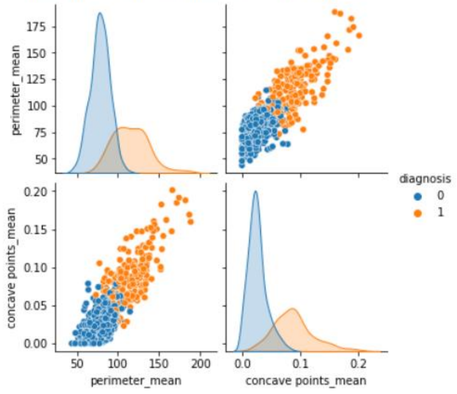
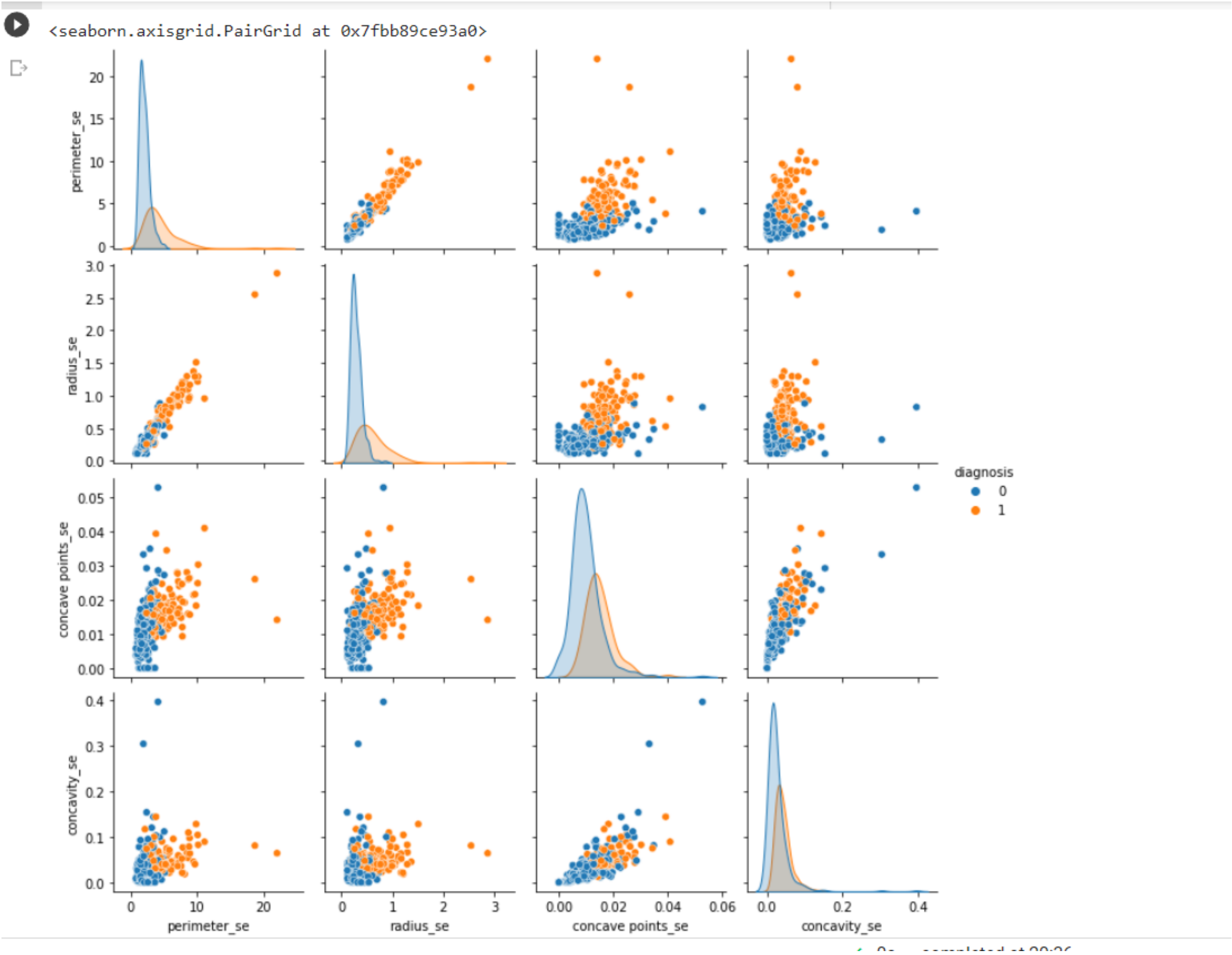


Figure 4: Pair plot of highly correlated values from first 10 attributes

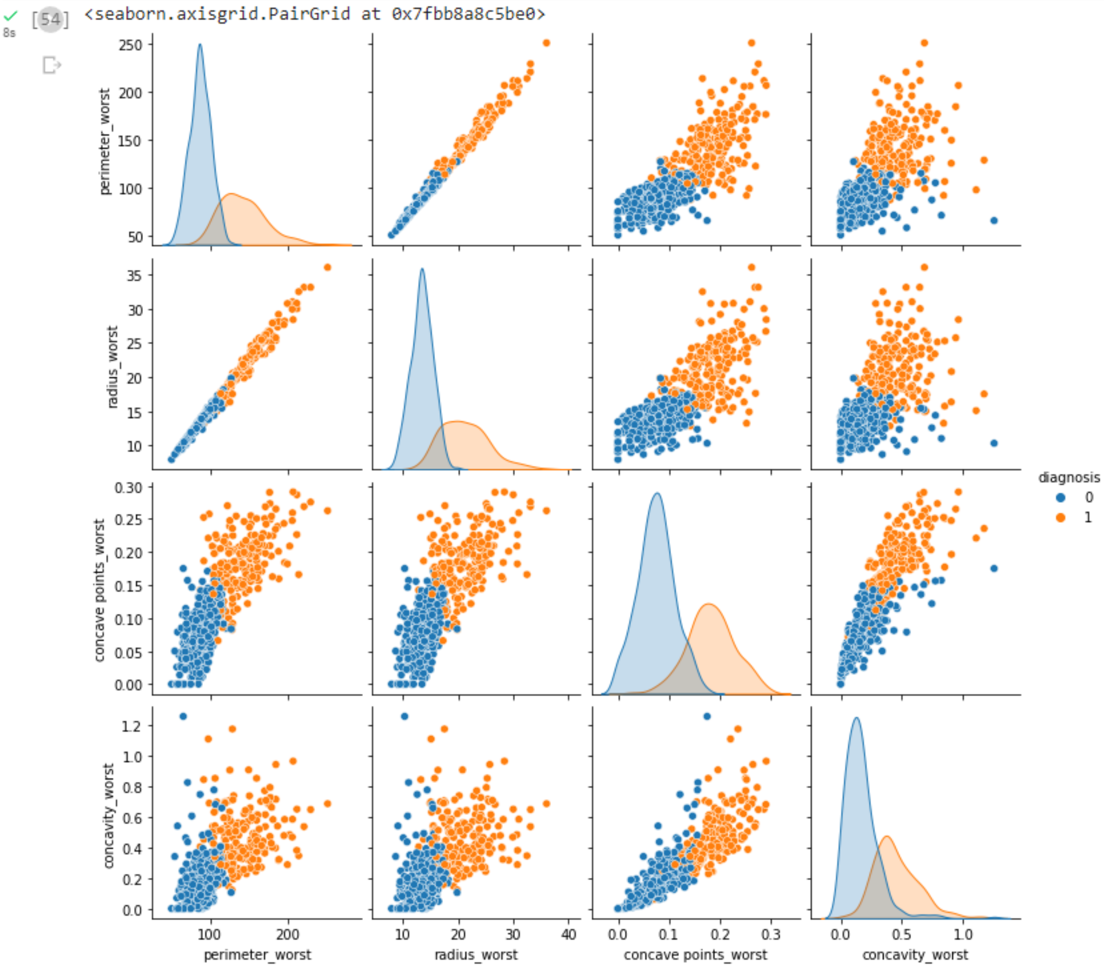
After comparing all the values, the columns which have highest correlation values among next 10 attributes are “perimeter_se”, “radius_se”, “concave points_se” and “concavity_se”. Hence, we have plotted a pair plot of these columns to check the malignant and benign data points, in order to crosscheck the correlation concept.



```
<seaborn.axisgrid.PairGrid at 0x7fbb89ce93a0>
```

Figure 5: Pair plot of highly correlated values from next 10 attributes

After comparing all the values, the columns which have highest correlation values among last 10 attributes are “perimeter_worst”, “radius_worst”,” concavity_worst” and “concave points_worst”. Hence, we have plotted a pair plot of these columns to check the malignant and benign data points, in order to crosscheck the correlation concept.



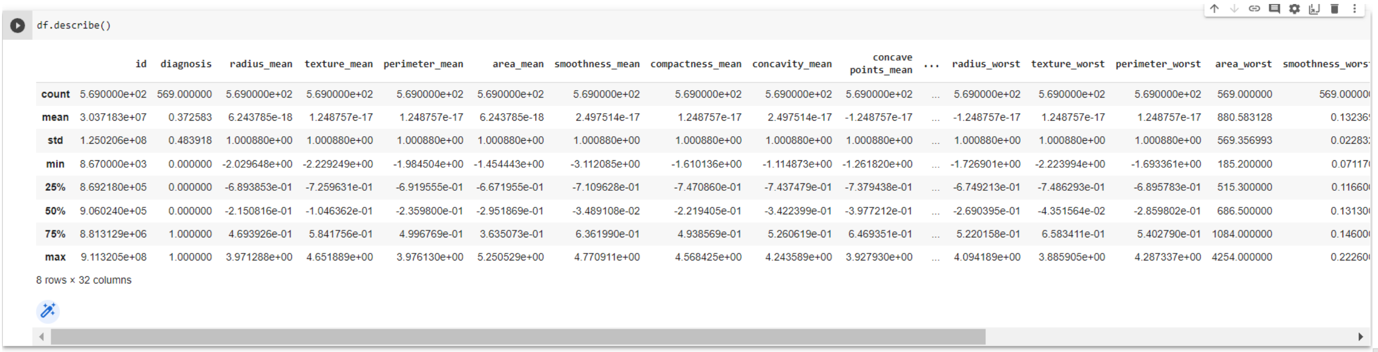
```
<seaborn.axisgrid.PairGrid at 0x7fbb8a8c5be0>
```

Figure 6: Pair plot of highly correlated values of last 10 attributes

Data Preprocessing:

Plotting box plots for the most important attributes contributing towards building a machine learning model to predict Breast cancer prediction.

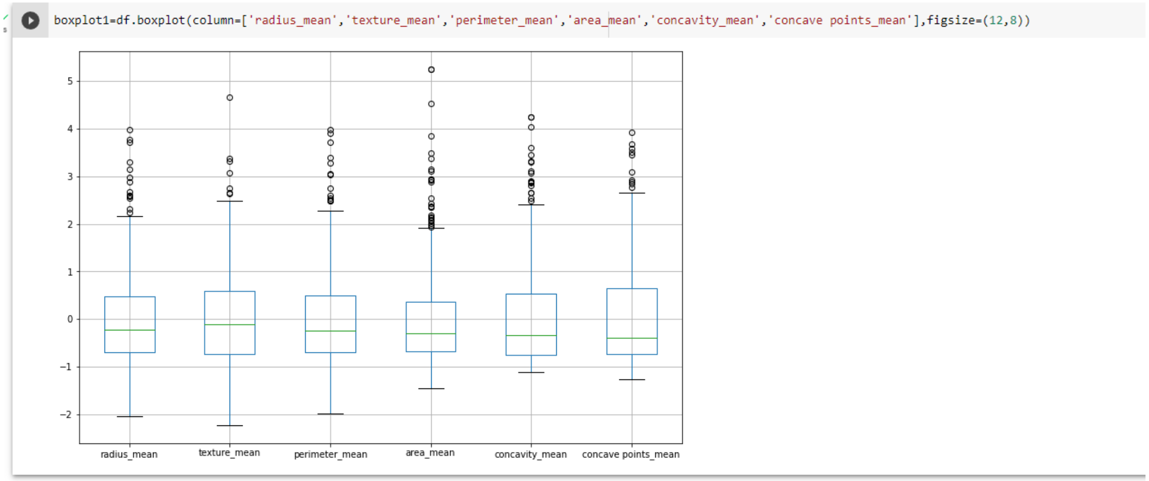
df.describe() – This function helps us understand the arithmetical attributions of the attributes present in the dataset.



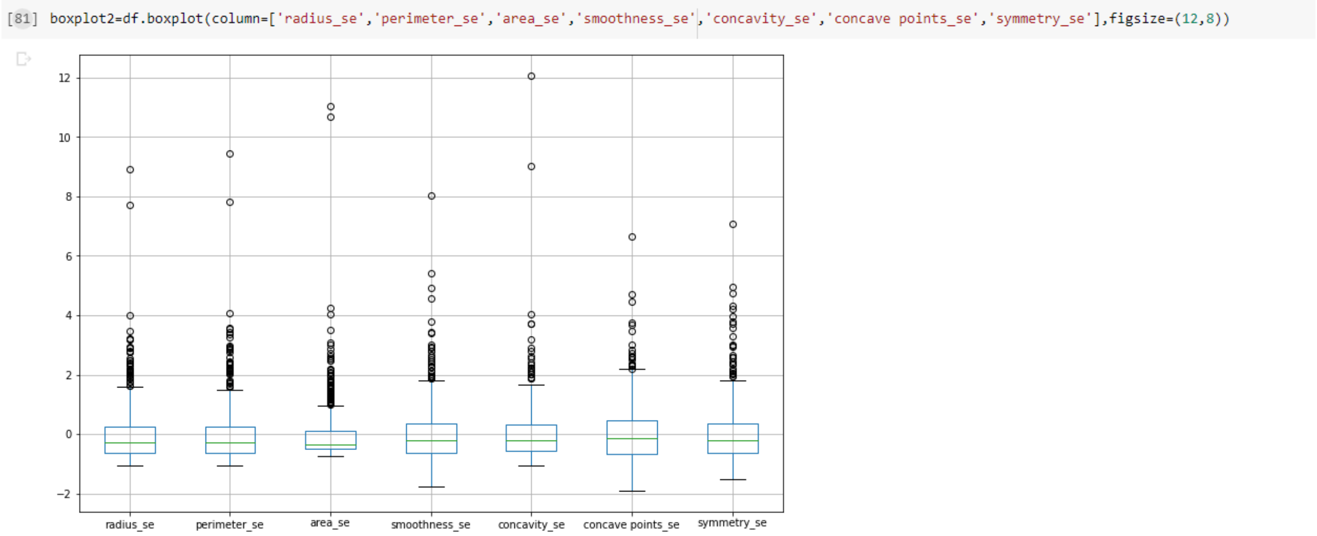
```
df.describe()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	radius_worst	texture_worst	perimeter_worst	area_worst	smoothness_wors
count	5.690000e+02	569.000000	5.690000e+02	5.690000e+02	5.690000e+02	5.690000e+02	5.690000e+02	5.690000e+02	5.690000e+02	5.690000e+02	...	5.690000e+02	5.690000e+02	5.690000e+02	569.000000	569.00000
mean	3.037183e+07	0.372583	6.243785e-18	1.248757e-17	1.248757e-17	6.243785e-18	2.497514e-17	1.248757e-17	2.497514e-17	-1.248757e-17	...	-1.248757e-17	1.248757e-17	1.248757e-17	880.583128	0.13236
std	1.250206e+08	0.483918	1.000880e+00	1.000880e+00	1.000880e+00	1.000880e+00	1.000880e+00	1.000880e+00	1.000880e+00	1.000880e+00	...	1.000880e+00	1.000880e+00	1.000880e+00	569.356993	0.02283
min	8.670000e+03	0.000000	-2.029648e+00	-2.229249e+00	-1.984504e+00	-1.454443e+00	-3.112085e+00	-1.610136e+00	-1.114873e+00	-1.261820e+00	...	-1.726901e+00	-2.223994e+00	-1.693361e+00	185.200000	0.07117
25%	8.692180e+05	0.000000	-6.893853e-01	-7.259631e-01	-6.919555e-01	-6.671955e-01	-7.109628e-01	-7.470860e-01	-7.437479e-01	-7.379438e-01	...	-6.749213e-01	-7.486293e-01	-6.895783e-01	515.300000	0.11660
50%	9.060240e+05	0.000000	-2.150816e-01	-1.046362e-01	-2.359800e-01	-2.951869e-01	-3.489108e-02	-2.219405e-01	-3.422399e-01	-3.977212e-01	...	-2.690395e-01	-4.351564e-02	-2.859802e-01	686.500000	0.13130
75%	8.813129e+06	1.000000	4.693926e-01	5.841756e-01	4.996769e-01	3.635073e-01	6.361990e-01	4.938569e-01	5.260619e-01	6.469351e-01	...	5.220158e-01	6.583411e-01	5.402790e-01	1084.000000	0.14600
max	9.113205e+08	1.000000	3.971288e+00	4.651889e+00	3.976130e+00	5.250529e+00	4.770911e+00	4.568425e+00	4.243589e+00	3.927930e+00	...	4.094189e+00	3.885905e+00	4.287337e+00	4254.000000	0.22260

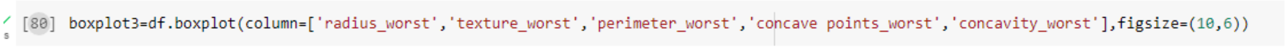
8 rows × 32 columns



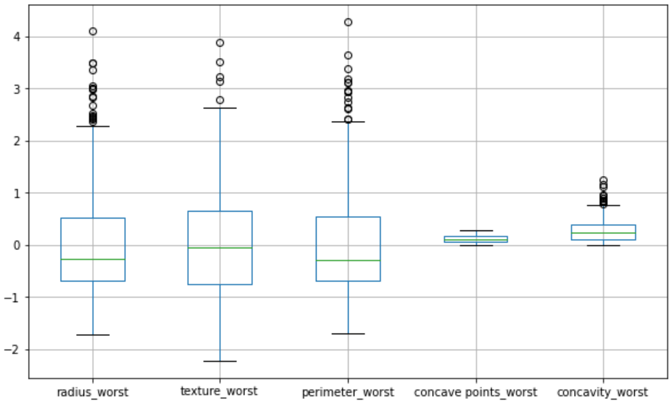
```
boxplot1=df.boxplot(column=['radius_mean','texture_mean','perimeter_mean','area_mean','concavity_mean','concave points_mean'],figsize=(12,8))
```



```
[81] boxplot2=df.boxplot(column=['radius_se','perimeter_se','area_se','smoothness_se','concavity_se','concave points_se','symmetry_se'],figsize=(12,8))
```

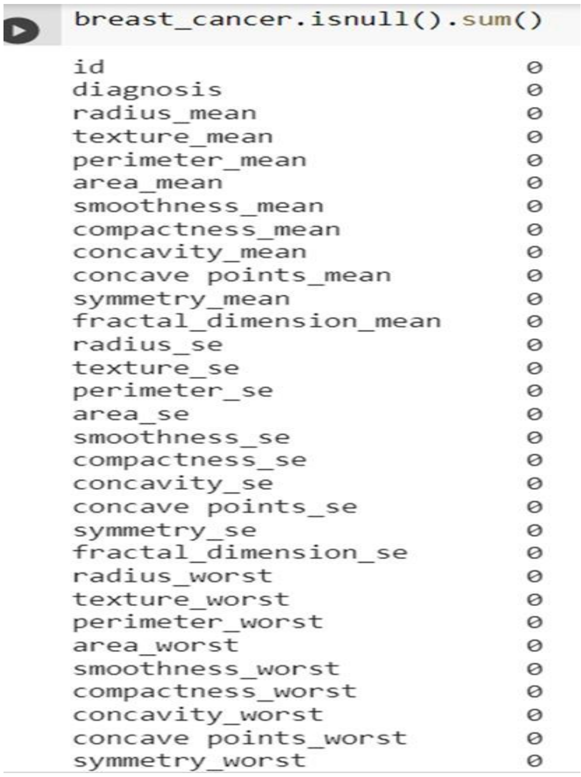


```
[80] boxplot3=df.boxplot(column=['radius_worst','texture_worst','perimeter_worst','concave points_worst','concavity_worst'],figsize=(10,6))
```



Data Cleaning

The breast cancer dataset consists of 30 predictors and one response variable and when we uploaded our dataset, we found no null values. So, we did not drop or impute any value to the attributes. Data cleaning is not necessary because our dataset is consistent, and no missing values and we are not standardizing the data and we checked the outliers.



```
breast_cancer.isnull().sum()

id                         0
diagnosis                  0
radius_mean                0
texture_mean               0
perimeter_mean             0
area_mean                  0
smoothness_mean            0
compactness_mean           0
concavity_mean             0
concave points_mean        0
symmetry_mean              0
fractal_dimension_mean     0
radius_se                  0
texture_se                 0
perimeter_se               0
area_se                    0
smoothness_se              0
compactness_se             0
concavity_se               0
concave points_se          0
symmetry_se                0
fractal_dimension_se       0
radius_worst               0
texture_worst              0
perimeter_worst            0
area_worst                 0
smoothness_worst           0
compactness_worst          0
concavity_worst            0
concave points_worst       0
symmetry_worst             0
```

Dimension Reduction:

Dimensionality reduction for a dataset depends on various factors such as the number of variables, the level of correlation among variables, the nature of the problem, the computational resources available, and the desired level of accuracy.

In the case of Breast cancer prediction dataset with 30 attributes, performing dimensionality reduction may not be necessary as the number of attributes is not very high. However, if some of the attributes are highly correlated, and there is a concern about overfitting, then performing dimensionality reduction can be beneficial.
Hence, we are not performing any dimensional reduction steps on the dataset, as all the attributes contribute to prediction of breast cancer.

Conclusion:

From the above observations we can find all the attributes that we selected for the feature selection are highly correlated and Perimeter mean, concave points mean, radius se, perimeter se, concave points se, concavity se, perimeter worst, radius worst, concave points worst, concavity worst are the attributes which are contributing more towards our prediction breast cancer. We can observe most of the data points to be malignant. We can conclude that the above selected attributes will play a major role for a better machine learning model.
Therefore, we conclude any predictive model developed using the dataset should consider role of these attributes and their relations.

Exploration of Candidate Data Mining Models, and Select the Final Model:

Breast cancer prediction is a common task in data mining and machine learning. In this task, we aim to predict whether a patient has breast cancer based on set of input features. To achieve this goal, we can explore and compare different data mining models and select the best based on its performance on the breast cancer prediction dataset.

Here are some common data mining models that can be used for breast cancer prediction:

1. Logistic Regression: A popular model for binary classification tasks, logistic regression models the probability of a patient having breast cancer given a set of input features.
2. Decision Trees: A tree-based model that recursively splits the input features based on their importance in predicting the target variable. Decision trees can be easily interpreted and visualized.
3. Random Forests: A type of ensemble model that combines multiple decision trees to improve performance and reduce overfitting.

4. Support Vector Machines (SVMs): A model that finds a hyperplane that separates the data into two classes with the largest margin. SVMs can be effective for high-dimensional datasets with a small number of observations.
5. Neural Networks: A complex model that uses multiple layers of nonlinear transformations to learn complex patterns in the data. Neural Networks can be effective for large datasets with many input features.

To select the best model for breast cancer prediction, we can follow these steps:

1. Load and preprocess the breast cancer prediction dataset, which may include steps such as data cleaning, normalization, and feature selection.
2. Split the dataset into training and test sets, typically using a ratio of 70-30 or 80-20.
3. Train each data mining model on the training set and evaluate its performance on the test set using metrics such as accuracy, precision, recall and F-1 score.
4. Compare the performance of each model and select the best one based on its overall performance and the specific needs of the task.

By comparing all the models, with breast cancer dataset, it is concluded that one model is most likely to predict with highest accuracy, i.e.;

1. **Logistic Regression:** It is a commonly used technique for binary classification tasks such as breast cancer prediction due to several reasons. First, logistic regression provides interpretable results in the form of coefficients that indicate the strength and direction of the relationship between each input feature and the predicted outcome. This can be important in medical applications where understanding the reasoning behind the predictions is necessary.

   Second, logistic regression is a relatively simple and fast algorithm that can be trained quickly on large datasets and is less prone to overfitting than some other models. Additionally, it is a robust algorithm that can handle noisy or incomplete data and is less sensitive to outliers than other models.

   Third, logistic regression has been found to be effective in several research studies for predicting breast cancer risk based on various features such as mammography results. Despite its simplicity, it can achieve high performance in many classification tasks including breast cancer prediction.

   Overall, logistic regression is a good choice for breast cancer prediction due its interpretability, low complexity, robustness and potential for high performance, with an accuracy of 0.94.

Model Performance Evaluation and Interpretation

Logistic regression is the classification model selected for the above breast cancer dataset. The accuracy for this model is 0.982. This is the highest accuracy from all the above models. The precision score of 0.977 and the misclassification is very low from the below confusion matrix. The ROC AUC score is 0.977 which implies the model is the perfect classifier. The highest specificity implies the model is performing well and the true positives are high and the classification is perfect according the below roc curve.
High specificity and sensitivity in a binary classification model imply that the model is able to effectively distinguish between positive and negative instances of the target class.

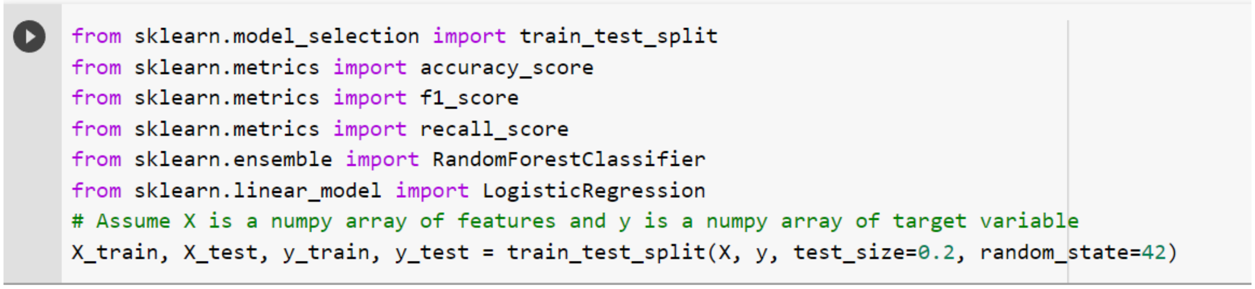
Specificity refers to the proportion of true negatives that are correctly identifies as negative by the model. A high specificity means that the model has a low false positive rate and is able to correctly identify most negative instances.

Sensitivity, also known as recall or true positive rate, refers to the proportion of true positives that are correctly identified as positive by the model. A high sensitivity means that the model has a low false negative rate and is able to correctly identify most positive instances.

In summary, a high accuracy, f1 score, high specificity and sensitivity indicates that the model has a low overall error rate and able to accurately classify both positive and negative instances of the target class. It is important to note that the optimal balance between specificity and sensitivity may vary depending on the specific application and the cost of false positives and false negatives.

Testing and checking which model is the best fit:

Explanation:



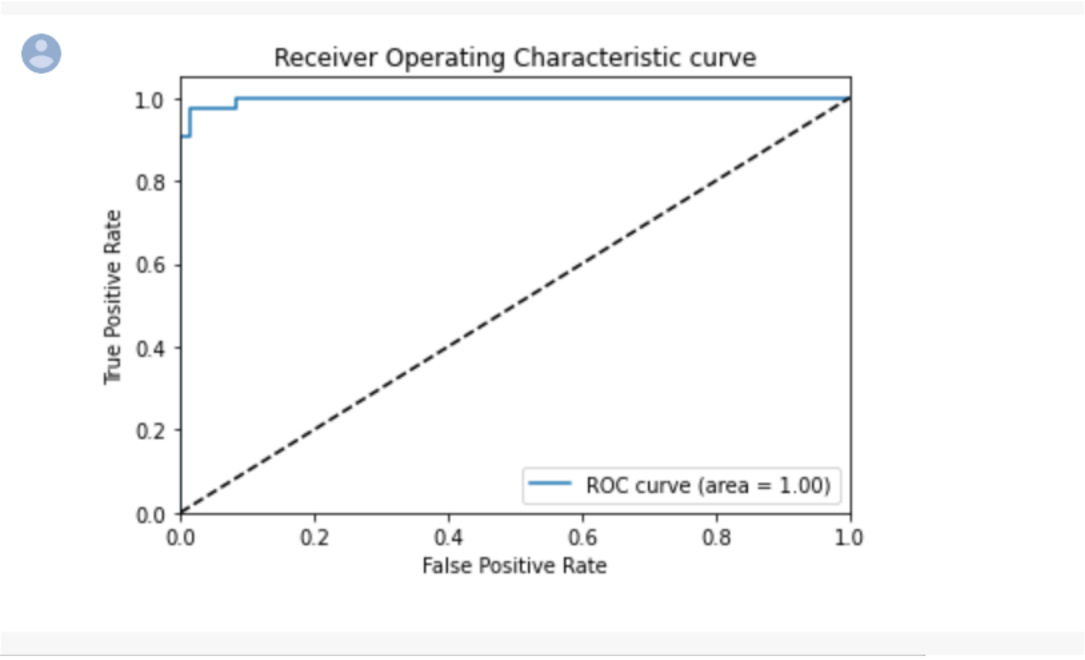
```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import f1_score
from sklearn.metrics import recall_score
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
# Assume X is a numpy array of features and y is a numpy array of target variable
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

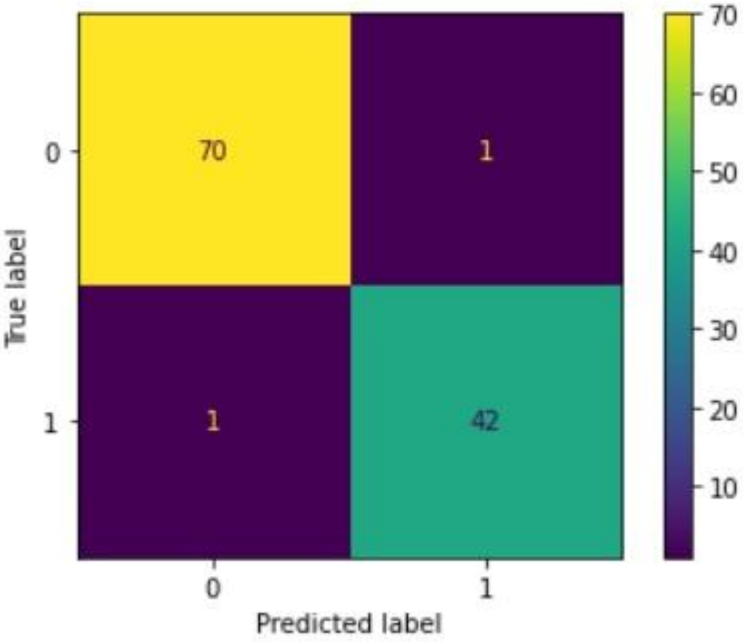
The above code is splitting the data into training and testing sets using the ‘train_test_split’ function from the ‘sklearn.model_selection’ module. This is a common practice in machine learning where a model is trained on a portion of the available data and then evaluated on the remaining unseen data.

The ‘x’ variable is assumed to contain the features, while ‘y’ is assumed to contain the target variable. The ‘test_size’ parameter specifies the proportion of the data to be used for testing, and ‘random_state’ ensures that the data is split in a reproducible way.

After splitting the data, the training set (‘x_train’ and ‘y_train’) is used to train various machine learning models, while the testing set (‘x_test’ and ‘y_test’) is used to evaluate the performance of these models.

1. **<u>Logistic Regression:</u>**
   The logistic regression model performed well on the breast cancer prediction data. The model achieved an accuracy of 0.982, indicating that it correctly classified 98.2% of the instances in the test set. The precision score of 0.977 indicates that the model was highly precise, correctly identifying 97.7% of the positive instances. The F1 score, which combines precision and recall, was also 0.977. The confusion matrix shows that the model had only 2 misclassifications: one false positive and one false negative. The sensitivity score of 0.977 indicates that the model was able to identify 97.7% of the positive instances, while the specificity score of 0.986 indicates that the model was able to correctly identify 98.6% of the negative instances. The ROC AUC score of 0.997 indicates that the model has good discriminative power, separating positive and negative instances with a high degree of accuracy. Overall, the logistic regression model is a promising candidate for breast cancer prediction, but we will need to compare its performance with other models to select the best one.



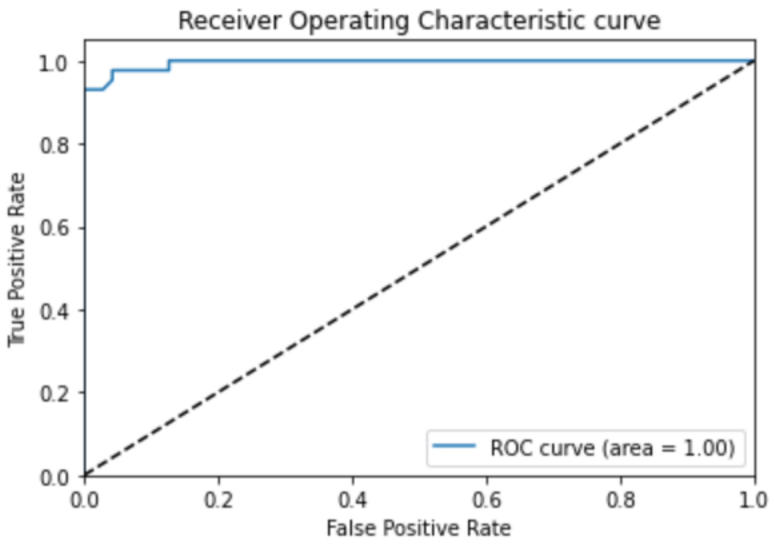


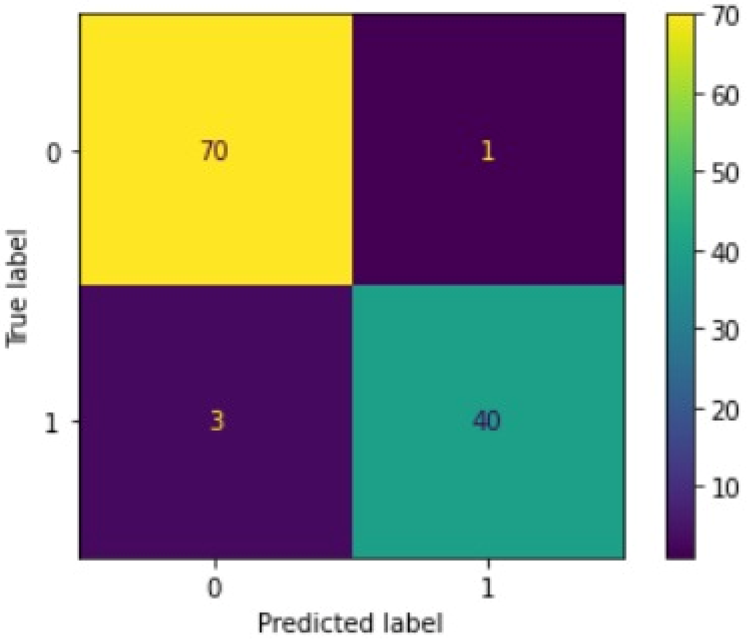
2. **<u>Random Forest Classifier:</u>**

The Random Forest Classifier was used to predict whether or not a patient has breast cancer based on various features. The model was trained using the Scikit-learn library in Python. After splitting the dataset into training and testing sets using the train_test_split method, the model was trained on the training set using the RandomForestClassifier method. The resulting model was then used to make predictions on the testing set.
The performance of the Random Forest Classifier was evaluated using several metrics including accuracy, precision, F1 score, sensitivity and specificity, and ROC AUC score.

The results of the evaluation showed that the Random Forest Classifier achieved an accuracy of 0.965, a precision of 0.976, and F1 score of 0.952, and a sensitivity of 0.93. The specificity was 0.986, and the ROC AUC score was 0.995.

Overall, the Random Forest Classifier performed well in predicting breast cancer based on the given features. The high accuracy, precision, and ROC AUC score indicate that the model is reliable in identifying patients with breast cancer. However, there is room for improvement in terms of the F1 score and sensitivity, which could be further optimized by turning the model's hyperparameters.



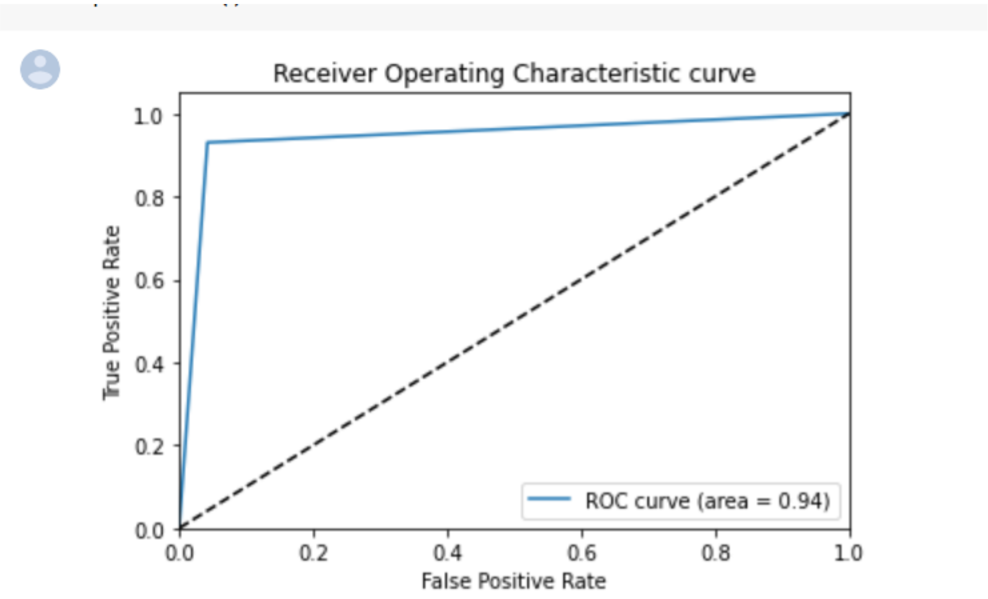


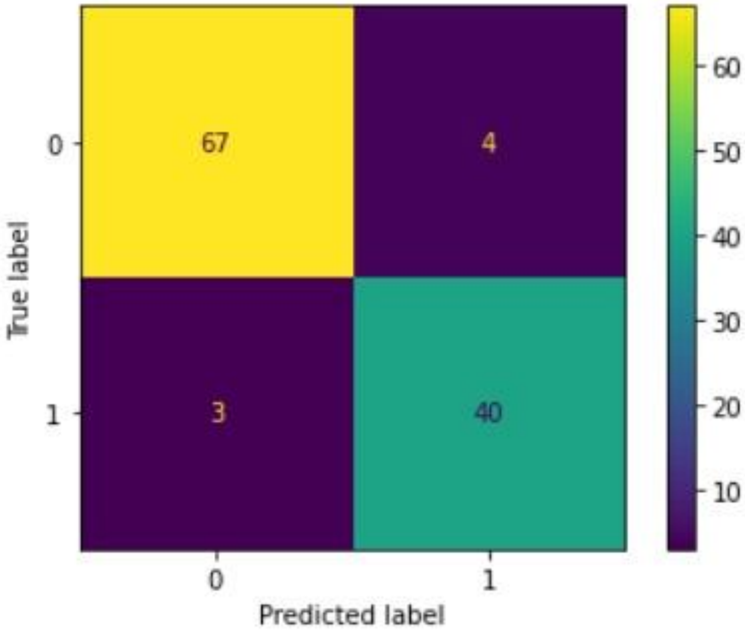
3. **<u>Decision Tree:</u>**

The Decision Tree model achieved an accuracy of 0.965, indicating that 96.5% of the test samples were correctly classified by the model. The precision score of 0.976 indicates that when the model predicted a sample as malignant, it was correct 97.6% of the time. The F1 score of 0.952 is a weighted average pf precision and recall, where a perfect score is 1.0 and the worst score is 0.0. The confusion matrix shows that the model correctly classified 70 benign and 40 malignant samples, with 1 false positive and 3 false negatives.

The sensitivity score of 0.93 indicates that the model correctly identified 93% of the malignant samples in the test set. The specificity score of 0.986 means that the model correctly identified 98.6% of the benign samples. A high specificity is desirable in this context since it indicates a low false positive rate.

Finally, the ROC AUC score of 0.944 is a measure of the model's ability to distinguish between the two classes. A score of 1.0 indicates perfect performance, while a score of 0.5 indicates good performance but is lower than that of the logistic regression and random forest models.



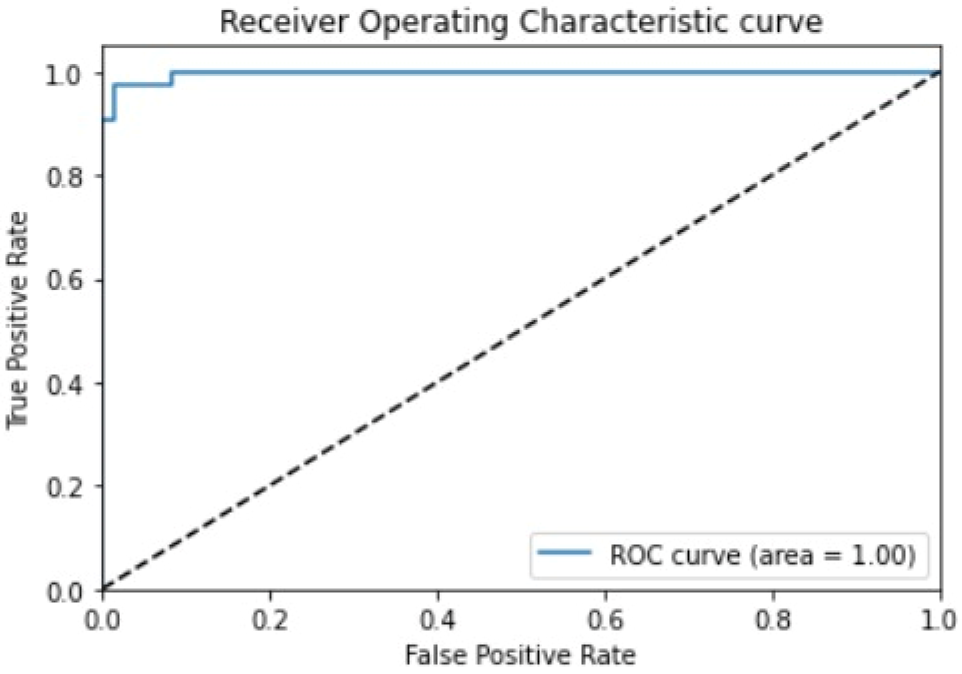


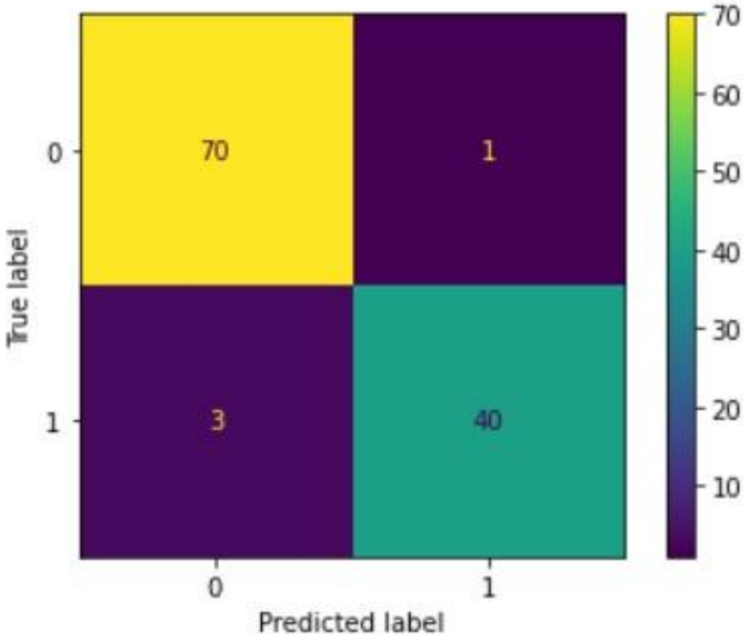
4. <u>Neural Networks:</u>

The accuracy score for the neural network classifier is 0.965, which means that it correctly classified 96.5% of the test samples. The precision score of 0.976 indicates that out of all the samples that the classifier predicted as positive, 97.6% were positive. The F1 score of 0.952 is a harmonic mean of precision and recall, and it measures the balance between precision and recall.

The confusion matrix shows the actual and predicted labels for the test set. In this case, there were 70 true negatives, 1 false positive, 3 false negatives, and 40 true positives. The sensitivity score of 0.93 indicates that the classifier correctly identified 93% of the positive samples, while specifically score of 0.986 indicates that it correctly identified 98.6% of the negative samples.

Finally, the ROC AUC score pf 0.977 indicates that the classifier has a high discrimination ability, i.e., it can distinguish between positive and negative samples with a high degree of accuracy.





Conclusion:

After evaluating all the machine learning models, it can be concluded that logistic regression performed the best on the given dataset. The model was able to accurately classify the instances with a high degree of accuracy and precision, as well as having a good balance between sensitivity and specificity. The remaining models also performed well but had slightly accurate scores and less balanced sensitivity and specificity compared to logistic regression.

Overall, logistic regression is the recommended model for this dataset.