

# **Breast Cancer Prediction**

Group 25

**Student 1:** Siva Vasanta Harika Mangu

**Student 2:** Raaga Sindhu Mangalagiri

[mangu.s@northeastern.edu](mailto:mangu.s@northeastern.edu)

[mangalagiri.r@northeastern.edu](mailto: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 Vasanta Harika Mangu

**Signature of Student 2:** Raaga Sindhu Mangalagiri

**Submission Date:** 04/21/2023

## **TABLE OF CONTENTS**

### **Contents**

1.Problem Statement: .....	3
2.Problem Definition: .....	3
3.Data Sources: .....	3
4.Data Description: .....	4
5.Insights .....	4
6.Exploratory Data Analysis .....	5
7.Data Preprocessing: .....	10
8.Data Cleaning .....	12
9.Dimension Reduction: .....	13
10.Exploration of Candidate Data Mining Models, and Select the Final Model:.....	14
11.Model Performance Evaluation and Interpretation.....	19
12.Project Results: .....	21
13.Project Impact .....	22

## **1.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.

## **2.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.

## **3.Data Sources:**

Kaggle

[Breast Cancer Dataset | Kaggle](#)

## **4.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 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.

## **5.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:  $\text{perimeter}^2/\text{area} - 1.0$
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  $\text{perimeter}^2/\text{area} - 1.0$
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  $\text{perimeter}^2/\text{area} - 1.0$
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.

## **6.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 cancer 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

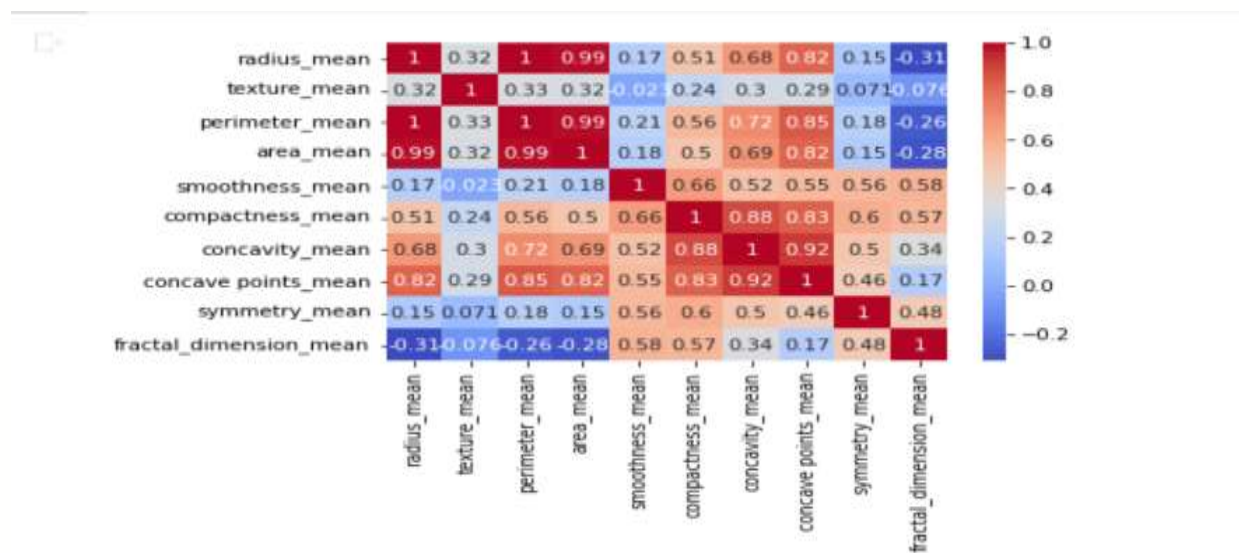


Figure1: Heatmap of correlation values of first 10 attributes

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.

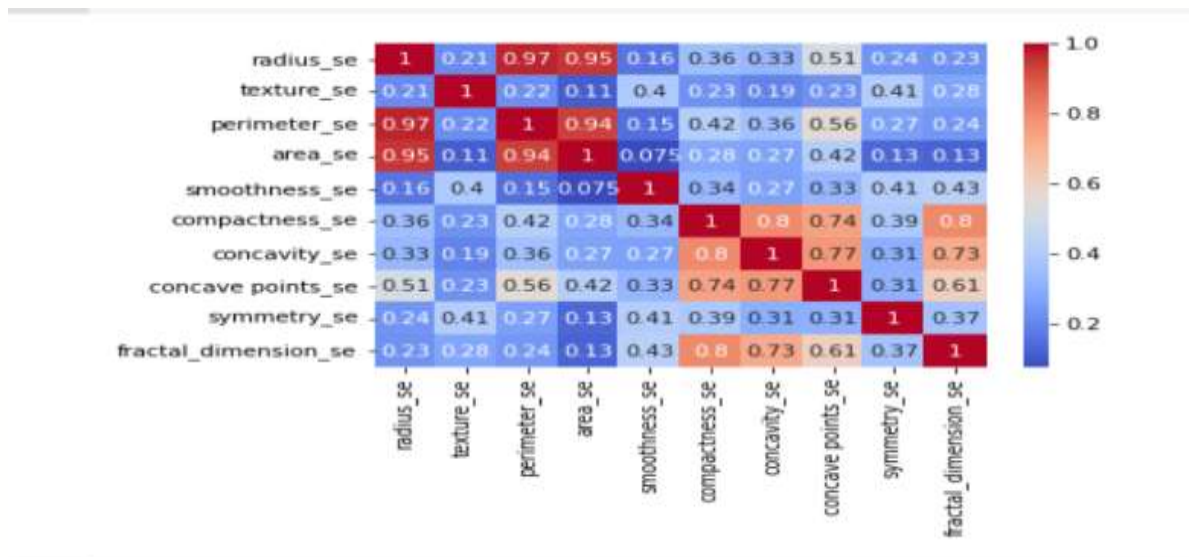


Figure 2: Heatmap of Correlation matrix of next 10 attributes

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.

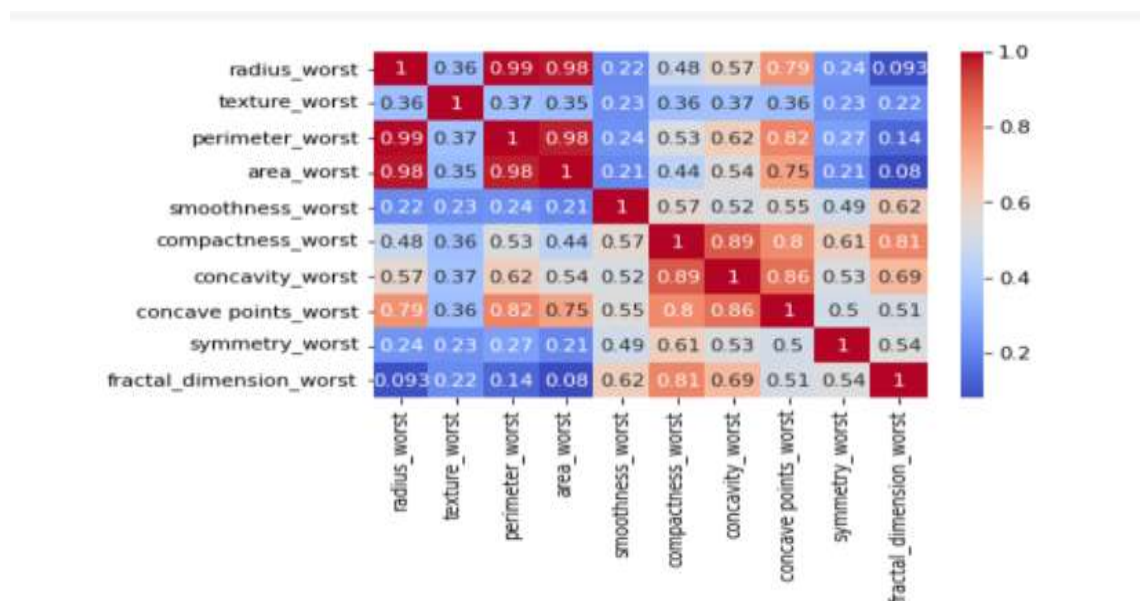


Figure 3: Heatmap of Correlation values of last 10 attributes

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.

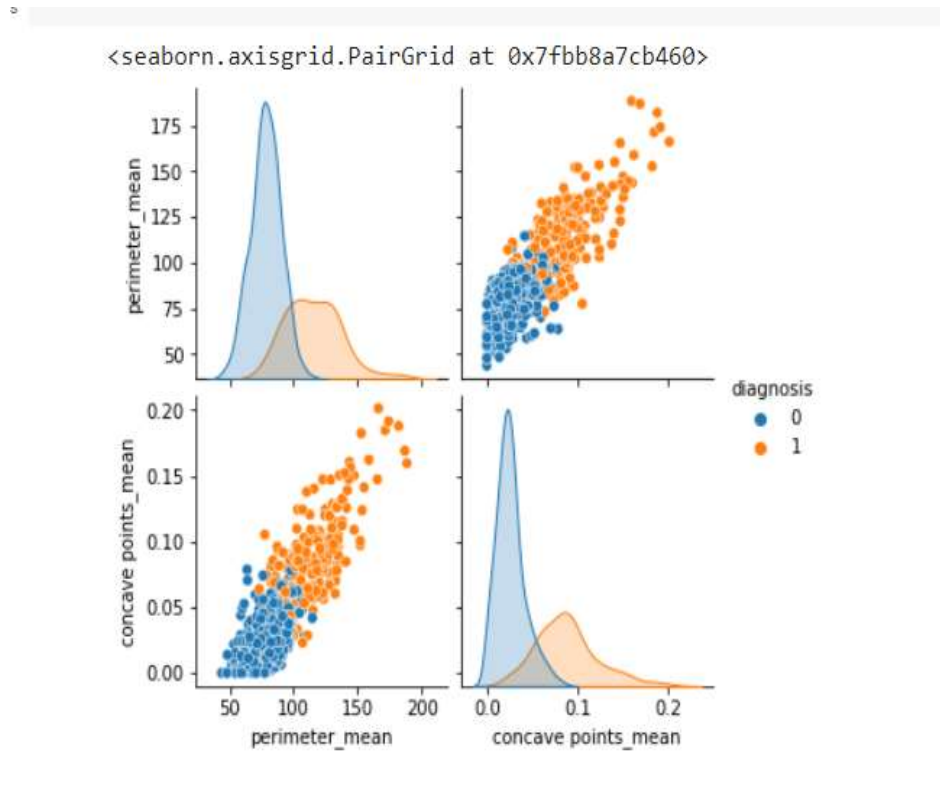


Figure 4: Pair plot of highly correlated values from first 10 attributes

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.



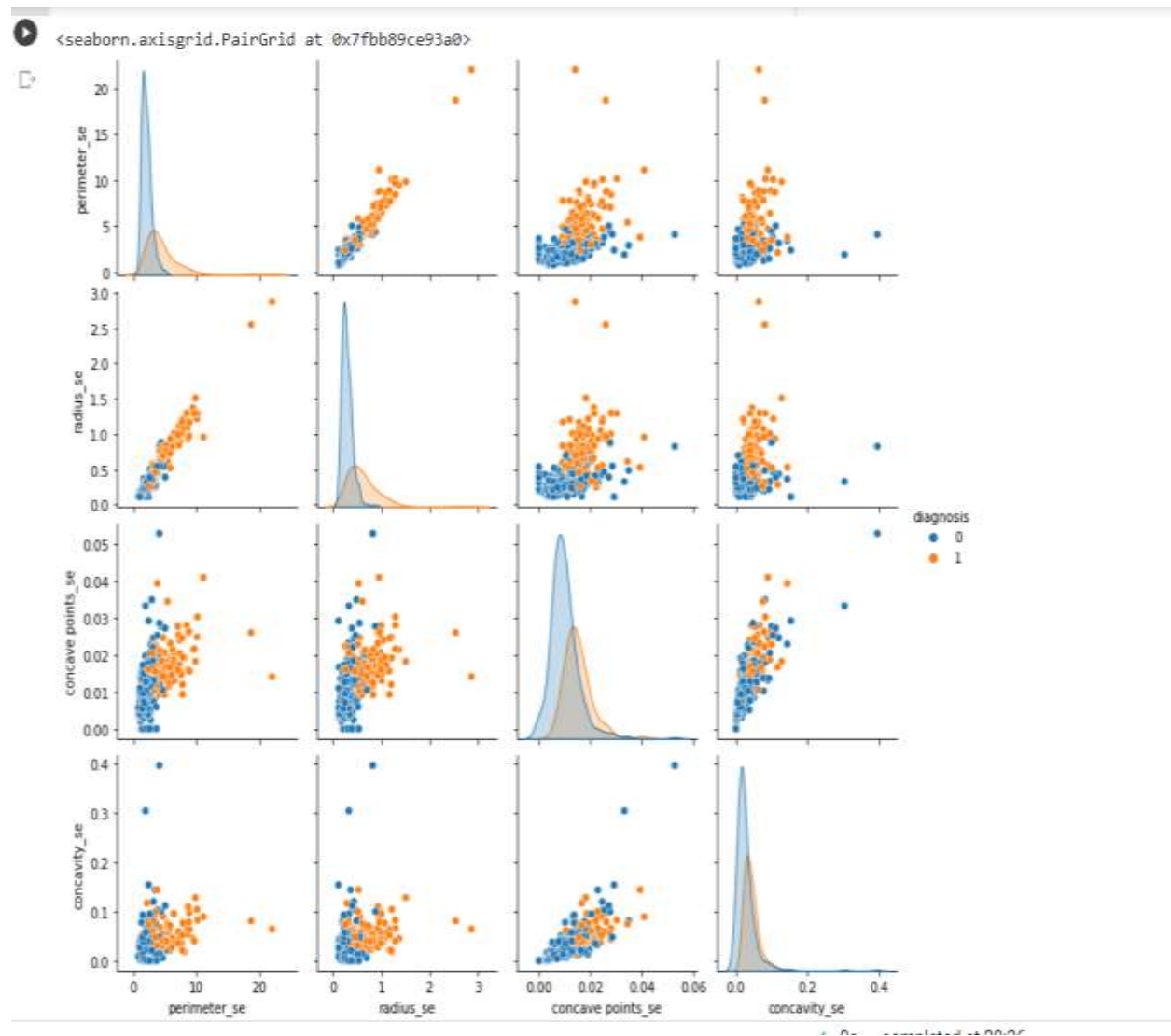


Figure 5: Pair plot of highly correlated values from next 10 attributes

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.

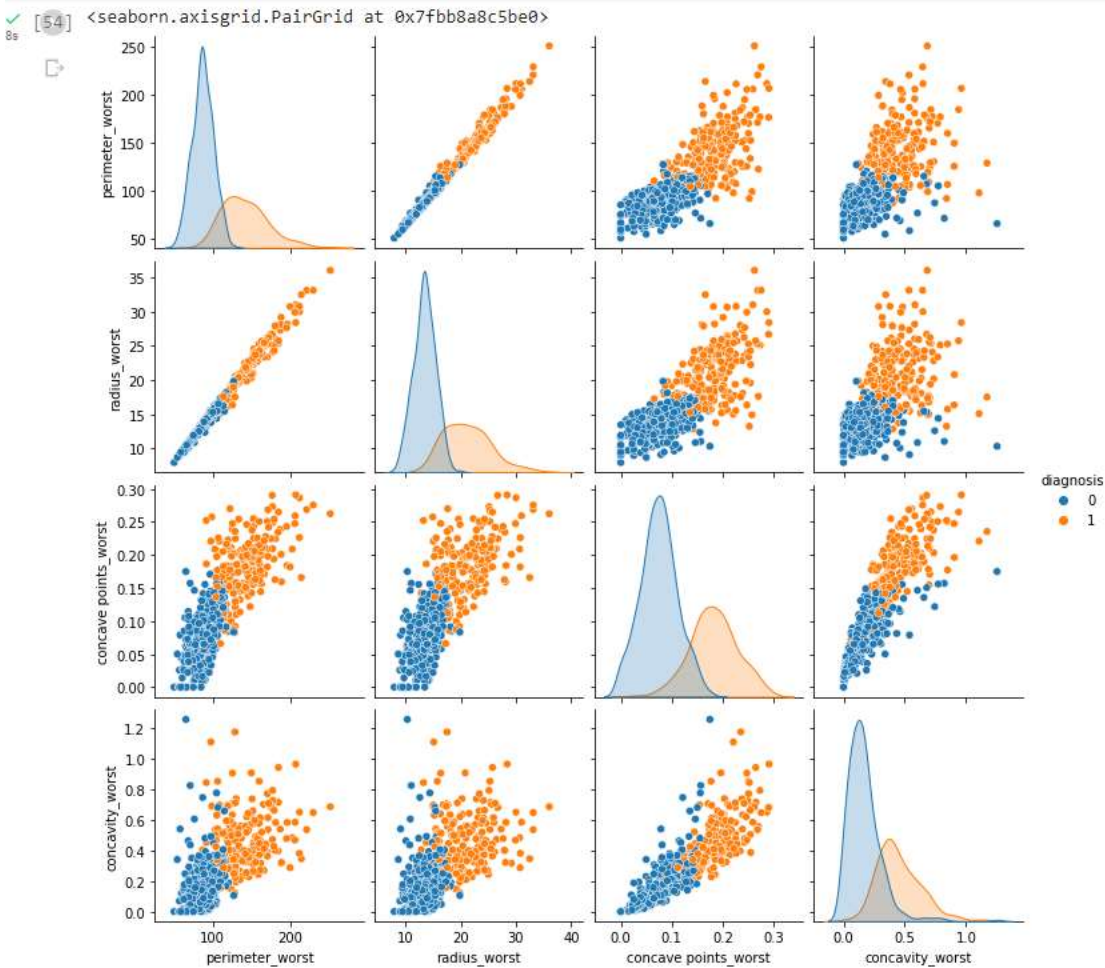


Figure 6: Pair plot of highly correlated values of last 10 attributes

## **7.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_worst
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	5.690000e+02	5.690000e+02
mean	3.037165e+07	0.372383	6.243705e-10	1.248757e-17	1.348757e-17	6.243705e-18	2.487514e-17	1.248757e-17	2.487514e-17	-1.248757e-17	...	-1.248757e-17	1.248757e-17	1.248757e-17	660.563128	0.132390
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	565.306993	0.022583
min	5.670000e+03	0.000000	-2.029648e+00	-2.229245e+00	-1.994504e+00	-1.434443e+00	-3.112385e+00	-1.610136e+00	-1.154873e+00	-1.261820e+00	...	-1.735901e+00	-2.229894e+00	-1.693361e+00	185.200000	0.071157
25%	5.682165e+05	0.000000	-6.893803e-01	-7.289631e-01	-6.918955e-01	-6.671955e-01	-7.108628e-01	-7.470860e-01	-7.437479e-01	-7.379438e-01	...	-6.745215e-01	-7.486295e-01	-6.896783e-01	515.300000	0.116609
50%	9.060240e+05	0.000000	-2.150816e-01	-1.045362e-01	-2.358800e-01	-2.951883e-01	-3.488108e-02	2.219405e-01	-3.422299e-01	-3.972125e-01	...	-2.698355e-01	-4.351564e-02	-2.856602e-01	686.500000	0.131339
75%	6.813129e+06	1.000000	4.893505e-01	5.841799e-01	4.199176e-01	3.835073e-01	6.381890e-01	4.908569e-01	5.268619e-01	6.488351e-01	...	5.220158e-01	6.583411e-01	5.402790e-01	1084.000000	0.146939
max	9.113205e+08	1.000000	5.971288e+00	4.651688e+00	3.976130e+00	5.253529e+00	4.770811e+00	4.568425e+00	4.240586e+00	3.827930e+00	...	4.084189e+00	3.885905e+00	4.287337e+00	4254.000000	0.222939

8 rows x 17 columns

Fig 7: Description of dataset

### Plotting to find the outliers:

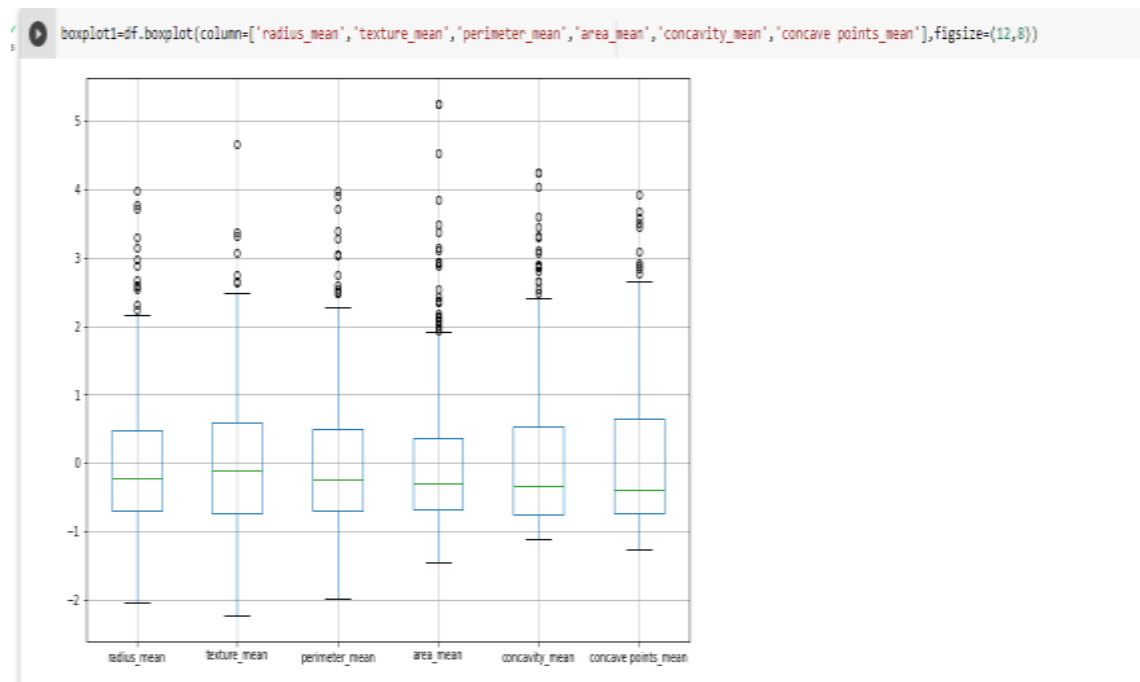


Fig8: Box plot for First set(mean) data

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

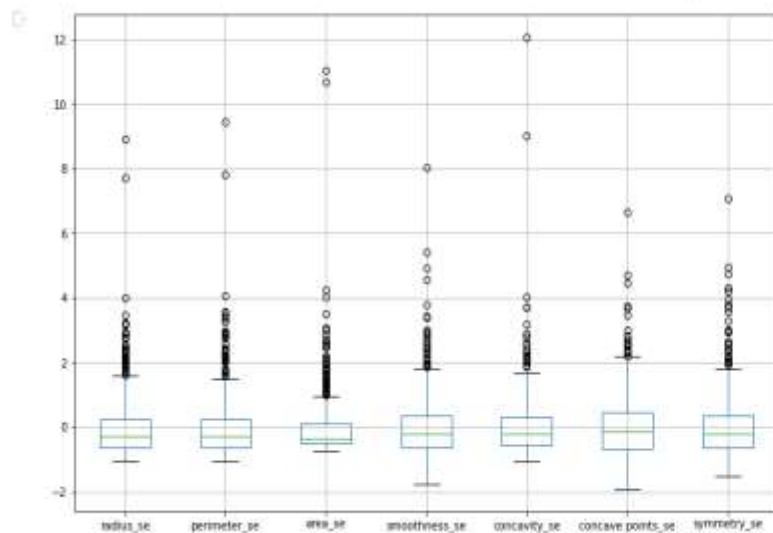


Fig 9:Box plot for second set (standard error)

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

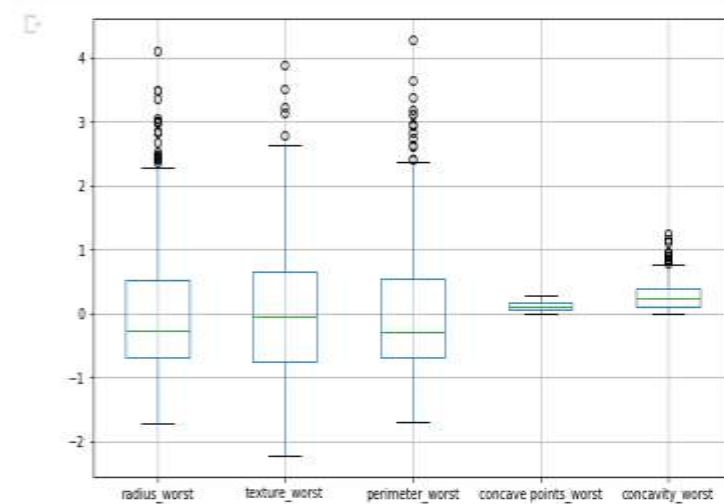
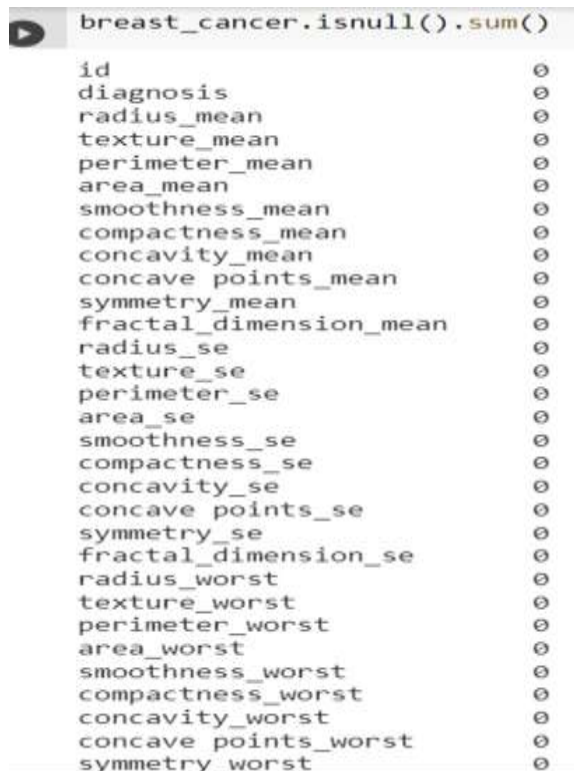


Fig 10:Box plot for third set (worst)

## 8. 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
```

Fig 11: Finding for null values (data cleaning)

## **9.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.

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.

## **10.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.

### **Feature selection:**

Pearson correlation is the technique used where we found the correlations between highly correlated variables and response variable.

List of parameters used for the data modelling are:

'id'

'diagnosis': response variable

'radius\_mean'

'texture\_mean'

'perimeter\_mean'

'area\_mean'

'smoothness\_mean'

'compactness\_mean'

'concavity\_mean'

'concave points\_mean'

'symmetry\_mean'

'fractal\_dimension\_mean'

'radius\_se'

'texture\_se'

'perimeter\_se'

'area\_se'

'smoothness\_se'

'compactness\_se'  
'concavity\_se'  
'concave points\_se'  
'symmetry\_se'  
'fractal\_dimension\_se'  
'radius\_worst'  
'texture\_worst'  
'perimeter\_worst'  
'area\_worst'  
'smoothness\_worst'  
'compactness\_worst'  
'concavity\_worst'  
'concave points\_worst'  
'symmetry\_worst'  
'fractal\_dimension\_worst'

### **SPLITTING DATASET:**

Splitting the data into training and testing in a proportion of 80% of training and 20% of testing  
We are performing random sampling where we reserved 80% of our data we considered as our training set and 20% of testing set using the function `sklearn.model selection.train_test split()` in python

### **MODEL SELECTION:**

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.
6. **Naïve Bayes:** A probabilistic model that can be used for classification tasks. It calculates the probability of each class based on input and selects class with highest probability as prediction. It is fast and known for simplicity and fast training time, making it a popular choice for many classification tasks.

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 two models are 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.

Performance metrics for logistic regression

	Accuracy	Precision	F1 score	Sensitivity	Specificity	ROC AUC score
<b>logistic_reg</b>	0.982	0.977	0.977	0.977	0.986	0.997

Fig 13: performance metrics for logistic regression

Confusion Matrix for logistic regression:

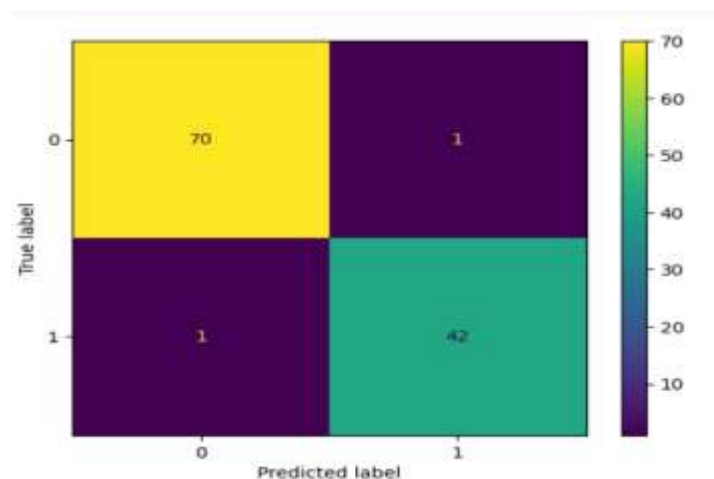


Fig 14: Confusion matrix for logistic regression

ROC Curve for logistic regression model:

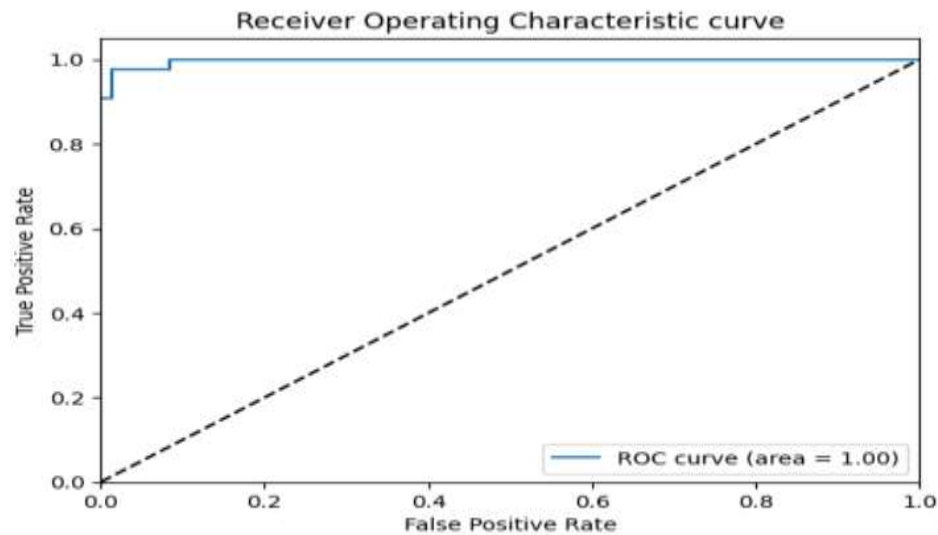


Fig 15: Roc curve for logistic regression

**2.Naïve Bayes:** Naive Bayes classifier is another commonly used technique for binary classification tasks such as breast cancer prediction, and it can be a good choice for several reasons. First, Naïve Bayes classifier is a simple and fast algorithm that can be trained quickly on large datasets. It is less computationally intensive than some other models such as logistic regression and decision trees.

Second, Naïve Bayes classifier performs well on high-dimensional data such as medical images and gene expression data, which can be important in breast cancer prediction. It is also robust to irrelevant features, meaning that it can still perform well even when there are many features that are not relevant to the classification task.

Performance metrics for Naïve Bayes

	Accuracy	Precision	F1 score	Sensitivity	Specificity	ROC AUC score
naive_bayes	0.965	0.976	0.952	0.93	0.986	0.997

Fig16: Performance metric for naïve bayes

## Confusion Matrix for naïve bayes

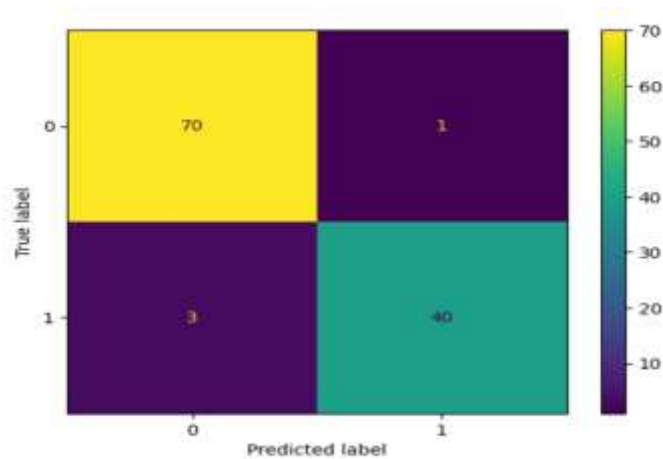


Fig 17: Confusion Matrix for Naïve bayes

## ROC curve for naïve bayes

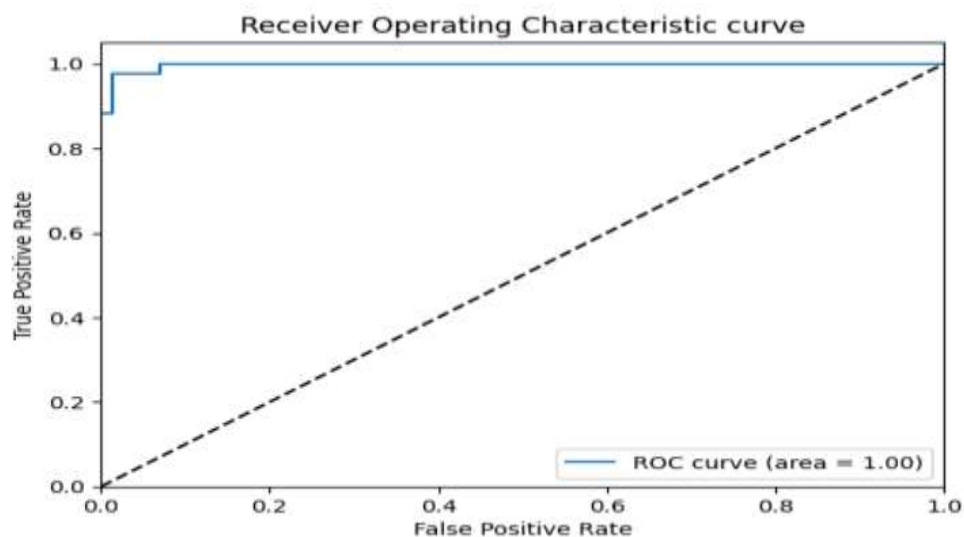


Fig 18: ROC Curve for Naïve bayes

**11. 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 to 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 identified 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 is 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.

#### Comparison between all the data models performed:

	Accuracy	Precision	F1 score	Sensitivity	Specificity	ROC AUC score
<b>logistic_reg</b>	0.982	0.977	0.977	0.977	0.986	0.997
<b>random_forest</b>	0.965	0.976	0.952	0.930	0.986	0.995
<b>naive_bayes</b>	0.965	0.976	0.952	0.930	0.986	0.997
<b>neural_networks</b>	0.974	0.955	0.966	0.977	0.972	0.997
<b>decision_tree</b>	0.930	0.907	0.907	0.907	0.944	0.925

Fig 19: Comparison between all data models

## **12.Project Results:**

The logistic regression is the best model among all other data models for breast cancer prediction. Breast cancer prediction is a vital task in medical research, and there are several project results that can be reported based on the analysis of breast cancer data.

Accuracy and performance of the predictive model: One of the primary results of a breast cancer prediction project is the accuracy and performance of the predictive model. This can be measured using metrics such as accuracy, precision, recall, F1 score, AUC-ROC, and lift chart. The project can report the best-performing model and the evaluation results of different models. The best model predicted for breast cancer prediction is logistic regression

1. Feature importance analysis: The breast cancer project can perform feature importance analysis to identify the most important features that contribute to breast cancer prediction. This can be done using methods such as correlation analysis, feature selection algorithms, and tree-based models. The project can report the most significant features and their importance scores. From our analysis we found the “perimeter\_se”, “radius\_se”, “concave points\_se”, “concavity\_se”, “perimeter\_mean”, “radius\_mean”, “concave points\_mean”, “concavity\_mean”, “perimeter\_worst”, “radius\_worst”, “concave points\_worst” and “concavity\_worst” are the most highly contributing factors for our model prediction.
2. Clinical implications: The project can discuss the clinical implications of breast cancer prediction, such as the potential for early diagnosis, personalized treatment, and improved patient outcomes. The project can also discuss the limitations of the predictive model and the need for further research.
3. Data preprocessing and cleaning: The project can report the data preprocessing and cleaning methods used, such as missing value imputation, outlier removal, and feature scaling. The project can also report any data quality issues and how they were addressed.
4. Visualization and interpretation: The project can use data visualization techniques to explore and interpret breast cancer data. We created plots such as histograms, scatterplots, and heatmaps to visualize the distribution and correlation of different features. The project can also use machine learning interpretation methods such as feature importance plots, correlation values, and decision trees to explain the predictions made by the model.

### **13. Project Impact**

The breast cancer prediction project helps us in early detection of cancer in women. It improves the chances of successful treatment and survival. The identification of high-risk individuals will be made easy. The statistical method used in our project is logistic regression as it allows a clear understanding of the relationship between predictors and response variable. The recall achieved during the model exploration is 0.977 which is the best score as the dataset we chose was a health care data and it is very important to find the recall i.e.; we are finding the true positive cases. The correct prediction helps us to recommend personalized medication to patients. Breast cancer prediction assists in optimizing resources in healthcare settings.

Overall, a breast cancer prediction project can provide valuable insights into the factors that contribute to breast cancer and how it can be predicted using machine learning models. These insights can help researchers and clinicians to develop better diagnostic and treatment strategies for breast cancer.