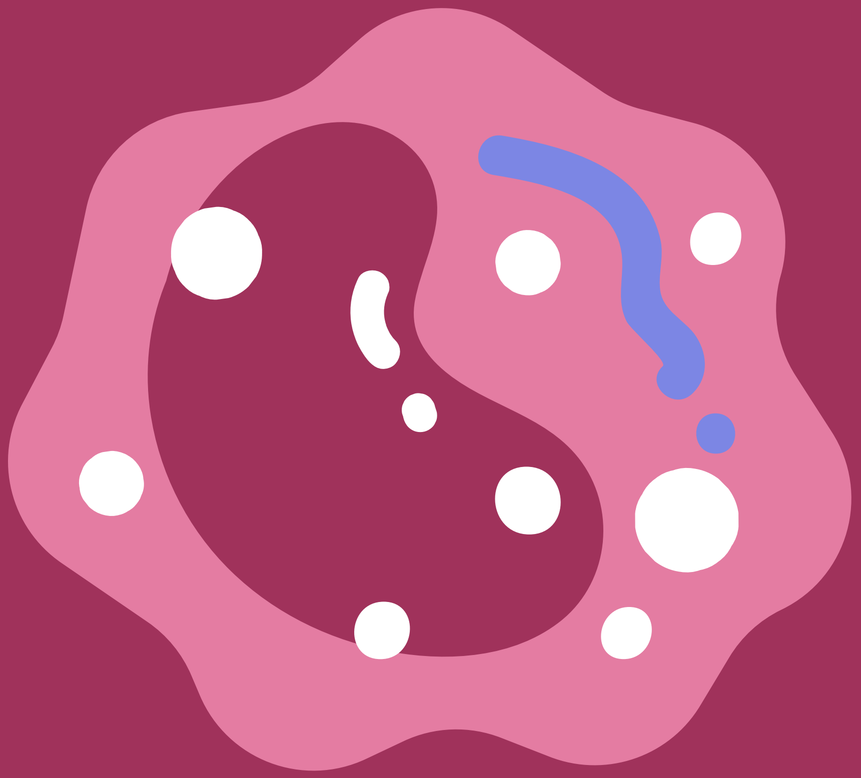


VARIANT ANALYSIS OF KNOWN BREAST CANCER GENES

GROUP MEMEBERS

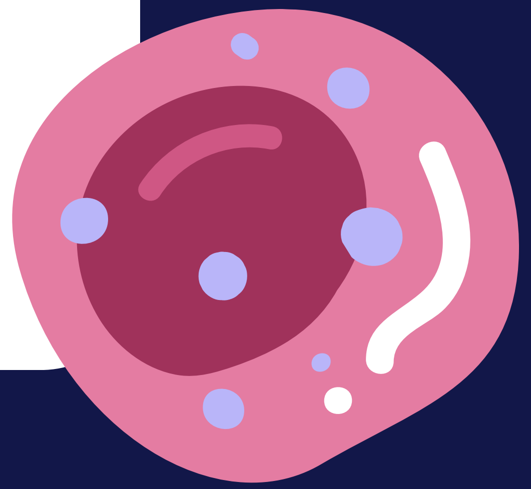
Aditya Kumar Sinha (2022034)
Nikhil Kumar (2022322)
Nikhil(2022321) Pandillapelly
Harshavardhini(2022345) Harsh
Vishwakarma(2022205)

INTRODUCTION

BREAST CANCER IS A COMPLEX DISEASE WITH VARIOUS GENETIC FACTORS CONTRIBUTING TO ITS DEVELOPMENT. WHILE NUMEROUS GENES HAVE BEEN IDENTIFIED AS POTENTIAL CONTRIBUTORS TO BREAST CANCER RISK, DETECTING VARIANTS WITHIN THESE GENES CAN PROVIDE VALUABLE INSIGHTS INTO THEIR ROLE IN DISEASE SUSCEPTIBILITY.

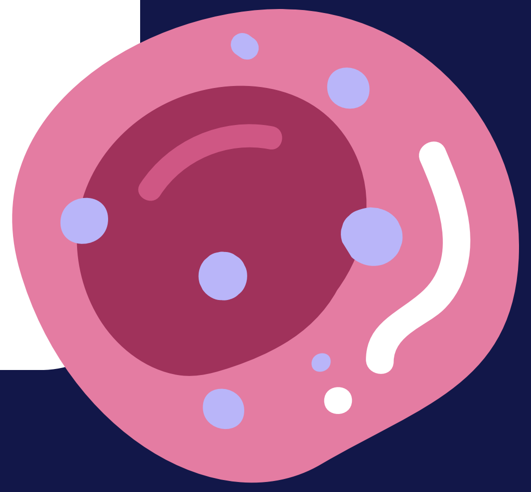
PROBLEM STATEMENT

The objective is to distinguish between breast cancer patients and healthy individuals with high precision and recall. The ultimate goal is to provide clinicians with a reliable tool for early diagnosis, enabling prompt intervention and personalized treatment strategies, thereby improving patient outcomes and reducing mortality rates associated with breast cancer.



SOLUTION

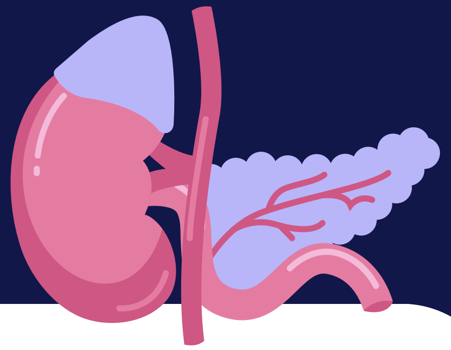
Developing a machine learning model for early breast cancer detection involves integrating diverse genomic, proteomic, and clinical datasets to extract informative features. The model aims to accurately classify individuals as either breast cancer patients or healthy individuals, utilizing advanced algorithms and robust evaluation methods and provides visualisation over data.



WORKFLOW

1. DATA ACQUISITION AND PREPROCESSING
2. MODEL TRAINING
3. MODEL INTERPRETATION
4. VALIDATION
5. DEPLOYMENT

PLAN OF ACTION



1) Data collection - Gather a comprehensive dataset comprising genomic data (gene expression profiles, mutations, copy number variations). Collect the hundreds of data of healthy person and breast cancer person.

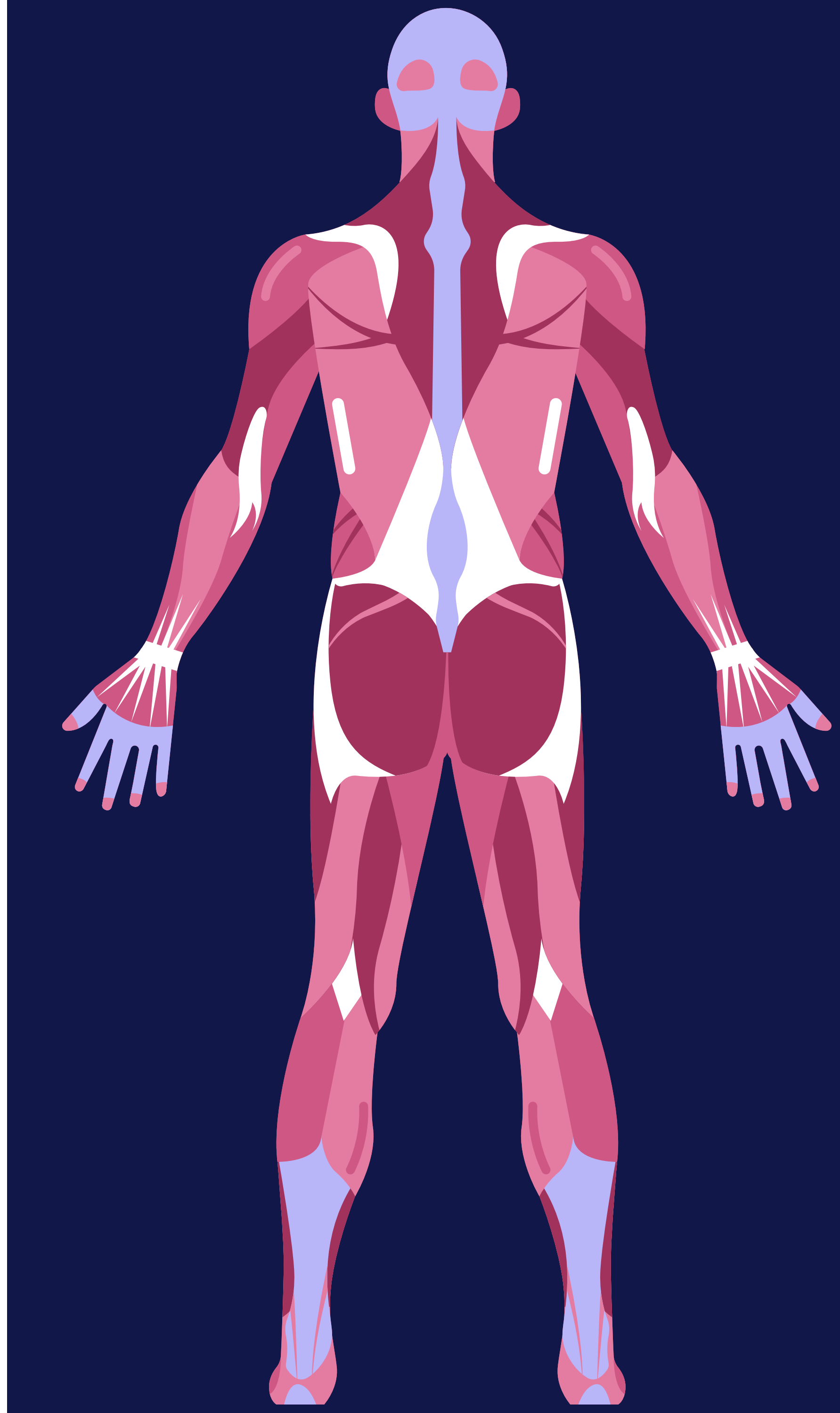
2) Data Preprocessing - Clean the dataset by normalizing features, removing any irrelevant or redundant features and converting the data into csv format for better readability.

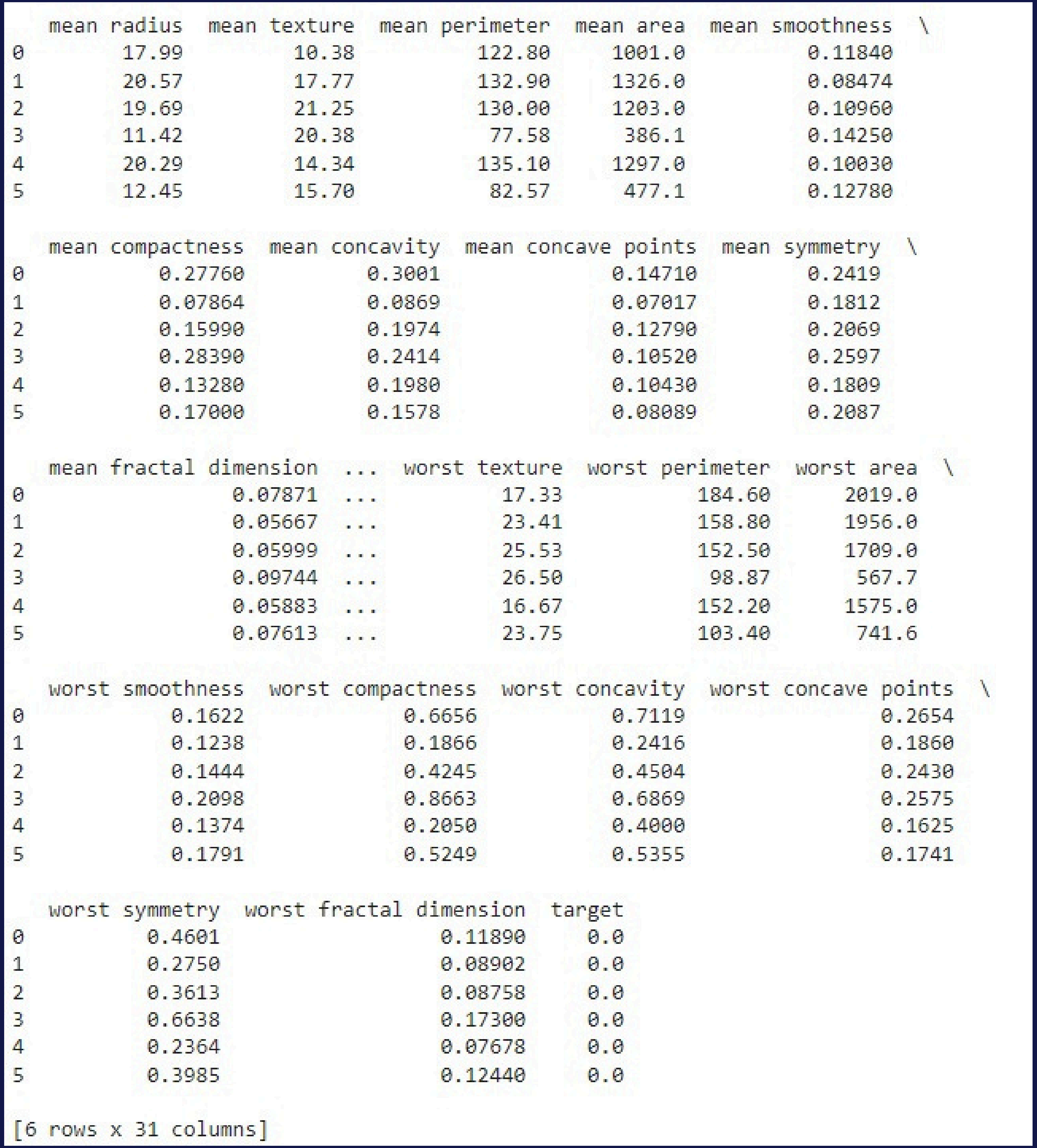
3) Model training - Split the dataset into training and testing sets. Train the selected machine learning models on the training data and evaluate their performance on the testing data.

4) Model Interpretation - Analyze the trained models to understand the importance of different features in predicting breast cancer compared to healthy. Visualize decision boundaries, feature importance, and model predictions to gain insights into the underlying biological mechanisms associated with breast cancer development.

5) Validation - various performance metrics such as accuracy, recall, and F1-score are calculated to assess the model's accuracy in classifying breast cancer patients and healthy individuals.

6) Implementation - involves putting the trained machine learning model into practical use, making it accessible to healthcare professionals etc. for real-world application in early breast cancer detection. Like pickling of data or converting it into binary format.

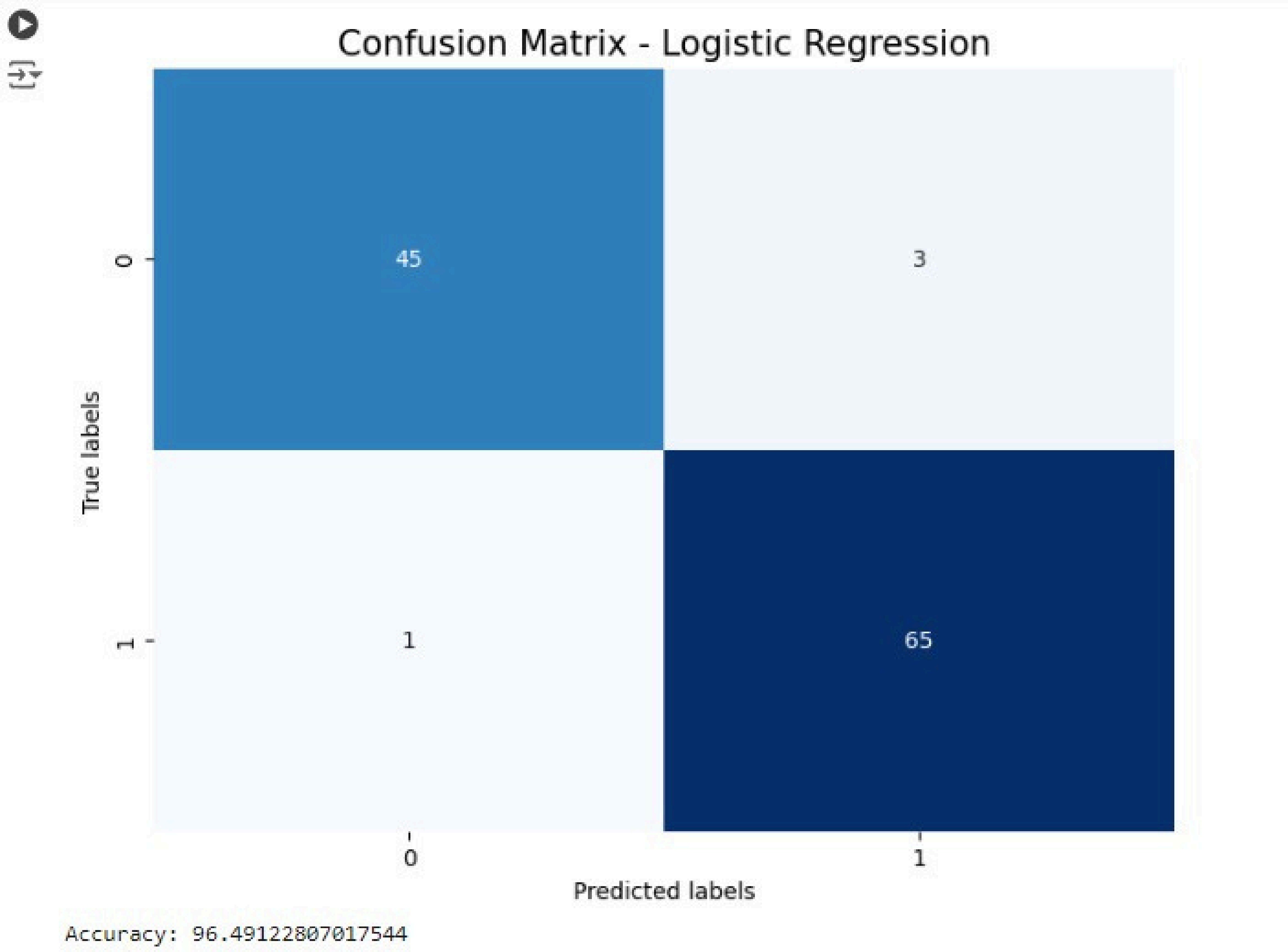




```
   mean radius  mean texture  mean perimeter  mean area  mean smoothness  \
0        17.99         10.38          122.80     1001.0          0.11840
1        20.57         17.77          132.90     1326.0          0.08474
2        19.69         21.25          130.00     1203.0          0.10960
3        11.42         20.38           77.58      386.1          0.14250
4        20.29         14.34          135.10     1297.0          0.10030
5        12.45         15.70           82.57      477.1          0.12780

   mean compactness  mean concavity  mean concave points  mean symmetry  \
0           0.27760          0.3001              0.14710         0.2419
1           0.07864          0.0869              0.07017         0.1812
2           0.15990          0.1974              0.12790         0.2069
3           0.28390          0.2414              0.10520         0.2597
4           0.13280          0.1980              0.10430         0.1809
5           0.17000          0.1578              0.08089         0.2087

   mean fractal dimension  ...  worst texture  worst perimeter  worst area  \
0                 0.07871  ...          17.33           184.60      2019.0
1                 0.05667  ...          23.41           158.80      1956.0
2                 0.05999  ...          25.53           152.50      1709.0
3                 0.09744  ...          26.50            98.87       567.7
4                 0.05883  ...          16.67           152.20      1575.0
5                 0.07613  ...          23.75           103.40       741.6

   worst smoothness  worst compactness  worst concavity  worst concave points  \
0            0.1622             0.6656           0.7119                0.2654
1            0.1238             0.1866           0.2416                0.1860
2            0.1444             0.4245           0.4504                0.2430
3            0.2098             0.8663           0.6869                0.2575
4            0.1374             0.2050           0.4000                0.1625
5            0.1791             0.5249           0.5355                0.1741

   worst symmetry  worst fractal dimension  target
0          0.4601                  0.11890     0.0
1          0.2750                  0.08902     0.0
2          0.3613                  0.08758     0.0
3          0.6638                  0.17300     0.0
4          0.2364                  0.07678     0.0
5          0.3985                  0.12440     0.0

[6 rows x 31 columns]
```

Data of 6 people containing 31 columns (attributes) like:

['mean radius' ,'mean texture', 'mean perimeter', 'mean area', 'mean smoothness', 'mean compactness' ,'mean concavity', 'mean concave points', 'mean symmetry' ,'mean fractal dimension', 'radius error' ,'texture error' , 'perimeter error', 'area error', 'smoothness error' ,'compactness error', 'concavity error', 'concave points error', 'symmetry error' ,'fractal dimension error', 'worst radius' ,'worst texture' ,'worst perimeter' , 'worst area', 'worst smoothness' ,'worst compactness', 'worst concavity', 'worst concave points', 'worst symmetry', 'worst fractal dimension']

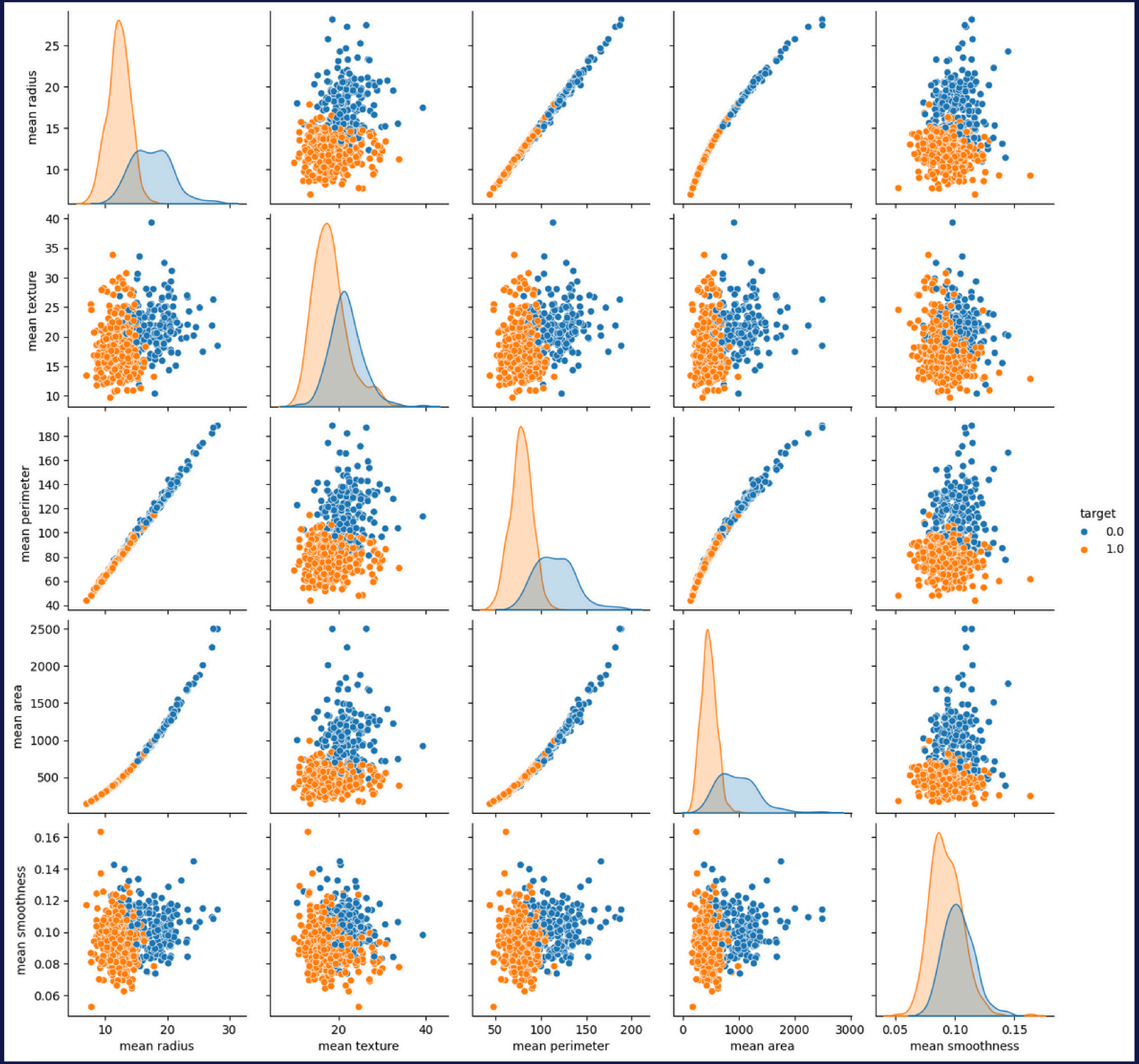


Confusion Matrix - Logistic Regression

	0	1
0	45	3
1	1	65

```
Accuracy: 96.49122807017544
```

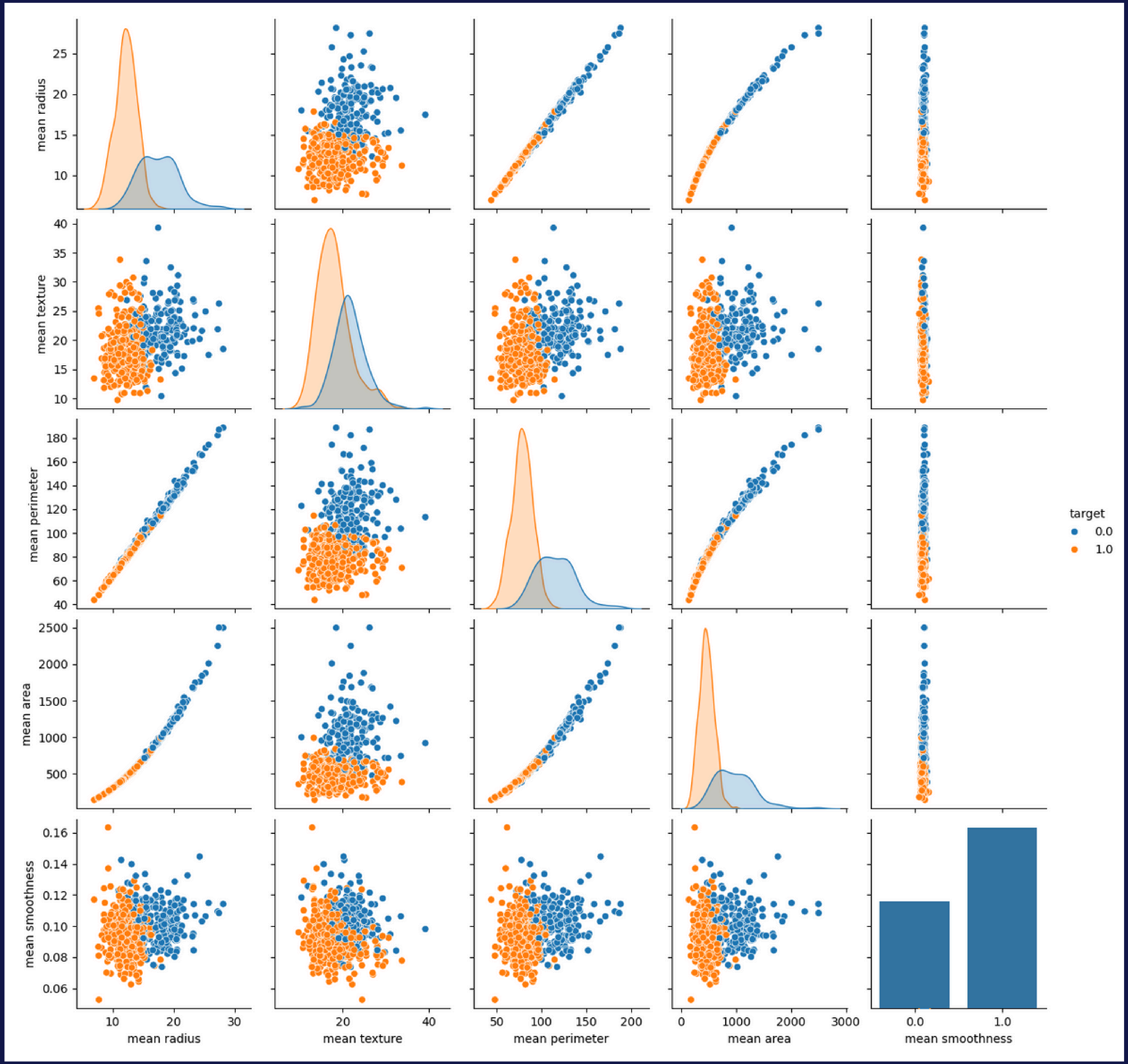
Confusion matrix and accuracy of trained model using logistic regression



PAIR PLOT

PEOPLE HAVING TUMOR

PEOPLE FREE OF TUMOR



COUNTER PLOT

PEOPLE HAVING TUMOR

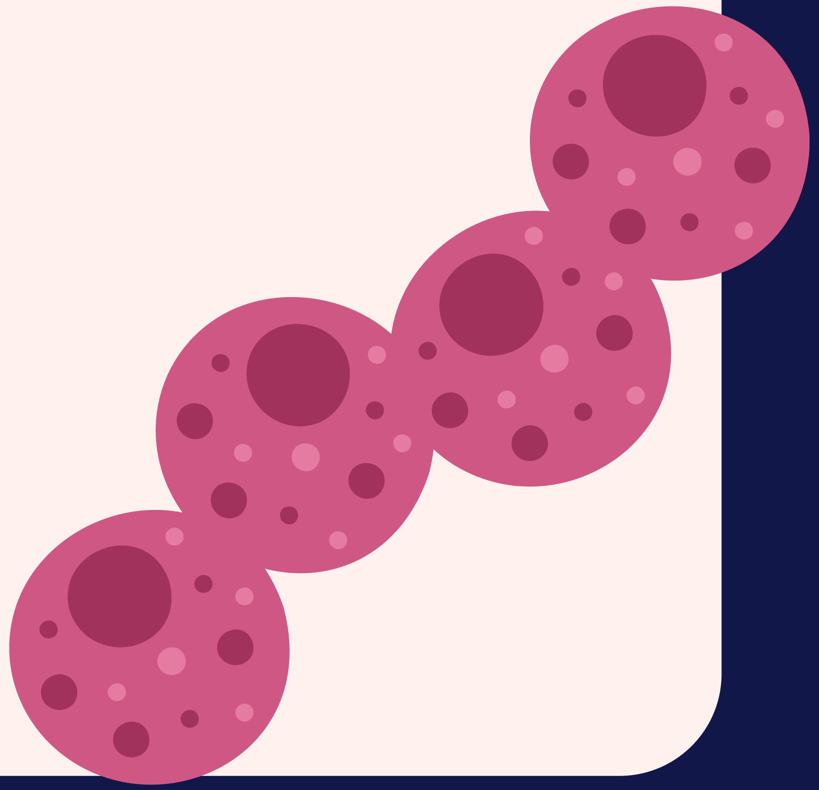
PEOPLE FREE OF TUMOR

REFERENCES:

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

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html

https://www.cancer.gov/ccg/research/genome-sequencing/tcga/studied-cancers/breast-lobular-carcinoma-study



CONTRIBUTIONS:

GROUP MEMEBERS

Aditya Kumar Sinha (2022034)-CODING WORK
Nikhil Kumar (2022322)- CODING WORK
Nikhil(2022321)- PPT MAKING
Pandillapelly Harshvardhini(2022345)- WORK FLOW
Harsh Vishwakarma(2022205)-DATA GATHERING