Breast Cancer Prediction

•••

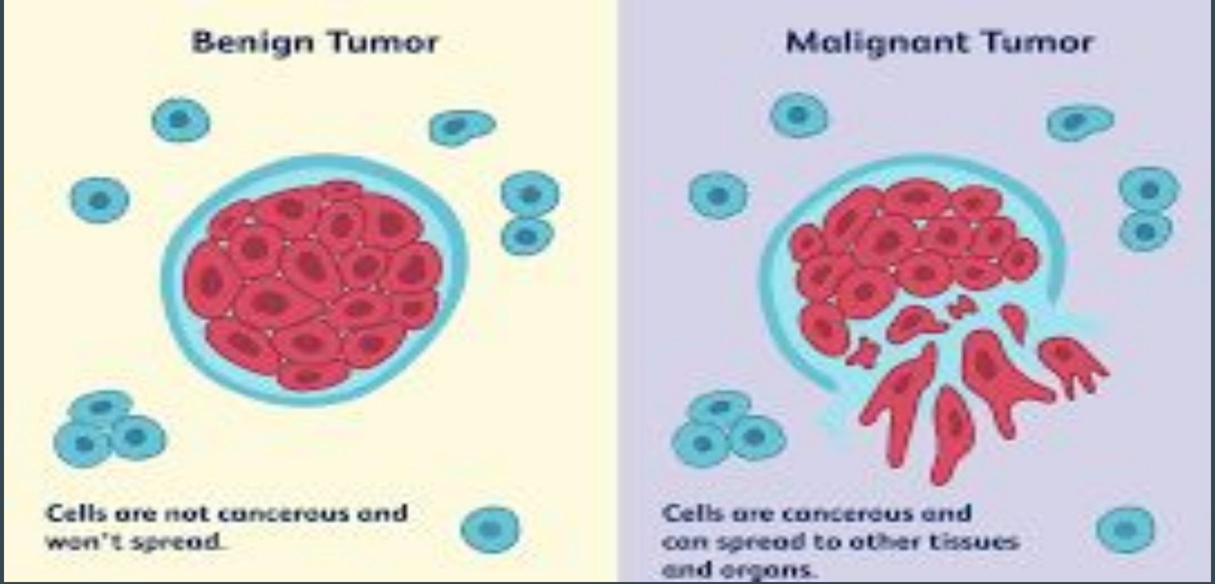
Name: Divya D Menon

OBJECTIVE:

- Predict the possibility of breast cancer among a wide range of data provided.
- Apply different ML algorithms to find the most accurate observation

EXPLANATION:

- Two types of tumors - Benign tumor and Malignant tumor


Benign Tumor
Malignant Tumor
Cells are not cancerous and won't spread.
Cells are cancerous and can spread to other tissues and organs.

DIFFERENCE:

BENIGN TUMOR:

- Non cancerous
- Capsulated
- Non invasive
- Slow growing
- Cells are normal
- Cells don't spread to other parts of the body

MALIGNANT TUMOR:

- Cancerous
- Non capsulated
- Fast growing
- Spread to other parts of the body
- Have abnormal shape
- Cells have large dark nuclei

DATA SET:

- Kaggle → Breast cancer Wisconsin data set
- Link → https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

STUDY OF DATASET:

1) ID number
2) Diagnosis (M = malignant, B = benign)

3) Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recorded with four significant digits.Missing attribute values: none

Class distribution: 357 benign, 212 malignant

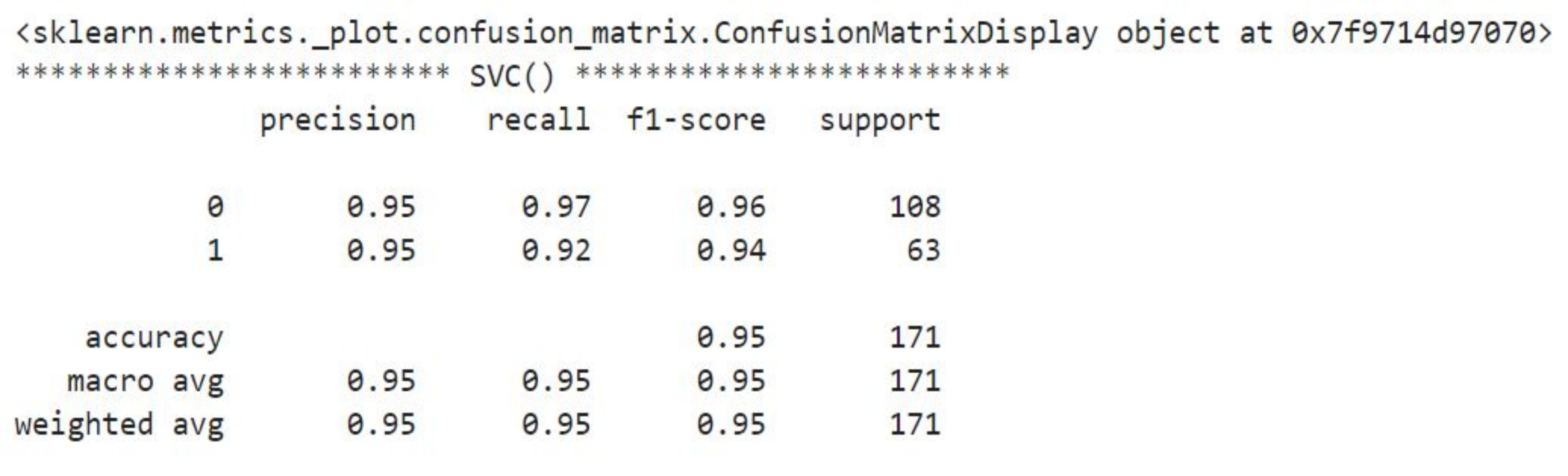
ALGORITHMS USED:

- KNN algorithm
- Naive Bayes algorithm
- Decision tree algorithm
- Random forest algorithm
- PCA

OBSERVATION:

- Analysis:https://colab.research.google.com/drive/13o51Q5qBW8IVZDwJmuWhmeCSSkMi5RKp#scrollTo=bKSCCsxySoA8
- On applying different algorithms the algorithm that is best fit is oversampled support Vector classification algorithm.

OBSERVATION cont..   (classification report)

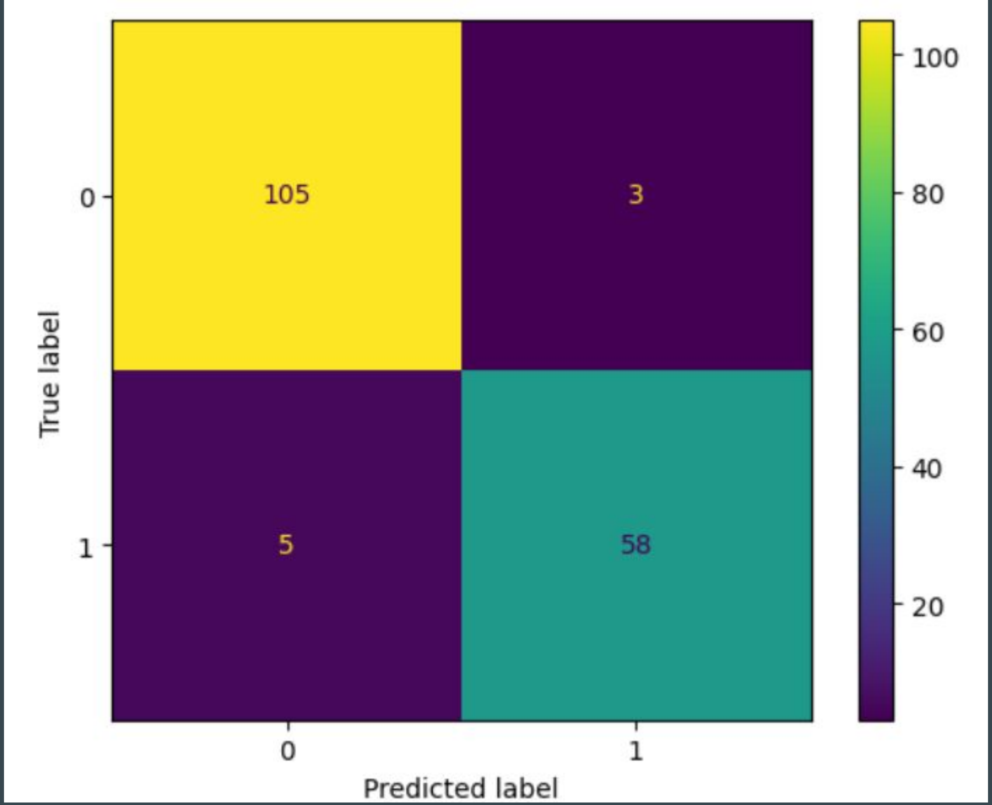


```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay object at 0x7f9714d97070>
************************* SVC() *************************
              precision    recall  f1-score   support

           0       0.95      0.97      0.96       108
           1       0.95      0.92      0.94        63

    accuracy                           0.95       171
   macro avg       0.95      0.95      0.95       171
weighted avg       0.95      0.95      0.95       171
```

OBSERVATION cont.... (confusion matrix)


True label
Predicted label