**DATA SCIENCE AND MACHINE LEARNING(DSML)**

**Ritvik Chawla**

**Reg No. – 209301056**

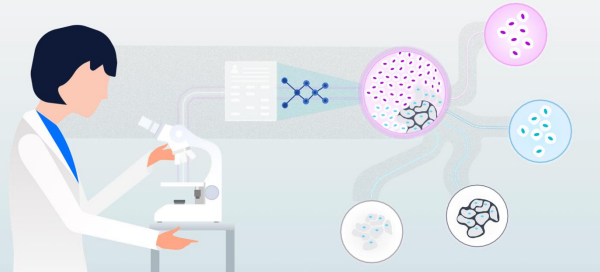
**Section E,**

**Submitted to – Mrs Shikha Mundra Ma’am**

**GITHUB LINK TO THE PROJECT:**

[**https://github.com/Ritvik10/DSML-Breast-Cancer-Detection-Project**](https://github.com/Ritvik10/DSML-Breast-Cancer-Detection-Project)

Breast Cancer Detection Using Machine Learning



**What is Breast Cancer?**

Cancer occurs when changes called mutations to take place in genes that regulate cell growth. The mutations let the cells divide and multiply in an uncontrolled, chaotic way. The cells keep on proliferating, producing copies that get progressively more abnormal. In most cases, the cell copies eventually end up forming a tumour. Breast cancer occurs when a malignant (cancerous) tumour originates in the breast. As breast cancer tumours mature, they may metastasize (spread) to other parts of the body. The primary route of metastasis is the lymphatic system which, ironically enough, is also the body's primary system for producing and transporting white blood cells and other cancer-fighting immune system cells throughout the body. Metastasized cancer cells that aren't destroyed by the lymphatic system's white blood cells move through the lymphatic vessels and settle in remote body locations, forming new tumours and perpetuating the disease process.

Breast cancer is not just a woman's disease. It is quite possible for men to get breast cancer, although it occurs less frequently in men than in women. Our discussion will focus primarily on breast cancer as it relates to women, but it should be noted that much of the information is also applicable to men.

**Facts And Figures**

Breast cancer is the most commonly occurring cancer in women and the second most common cancer overall. There were over 2 million new cases in 2018.

**Prevalence**

1.Asia

Percentage of world population: 59 Percentage of new breast cancer cases: 39 Percentage of breast cancer deaths: 44

2. Africa

Percentage of world population: 15 Percentage of new breast cancer cases: 8 Percentage of breast cancer deaths: 12

3. U.S. and Canada

Percentage of world population: 5 Percentage of new breast cancer cases: 15 Percentage of breast cancer deaths: 9

(Data from Global Cancer Facts and Figures, 3rd Edition, page 37)

**Incidence rates per 100,000 women**

1. Countries with the highest incidence: The Netherlands: 95.3 France: 94.6 U.S: (white people only - other races have lower incidence): 90.6

2. Countries with the lowest incidence:

Thailand: 25.6 Algeria: 29.8 India: 30.9

(Data from Global Cancer Facts and Figures, 3rd Edition, page 42)

The American Cancer Society's estimates for breast cancer in the United States for 2019 are:

• About 268,600 new cases of invasive breast cancer will be diagnosed in women.

• About 62,930 new carcinoma cases in situ (CIS) cases will be diagnosed (CIS is non-invasive and the earliest form of breast cancer).

• About 41,760 women will die from breast cancer.

**Role Of Machine Learning In the Detection Of Breast Cancer**

A mammogram is an x-ray picture of the breast. It can be used to check for breast cancer in women who have no signs or symptoms of the disease. It can also be used if you have a lump or other sign of breast cancer.

Screening mammography is the type of mammogram that checks you when you have no symptoms. It can help reduce the number of deaths from breast cancer among women ages 40 to 70. But it can also have drawbacks. Mammograms can sometimes find something that looks abnormal but isn't cancer. This leads to further testing and can cause you anxiety. Sometimes mammograms can miss cancer when it is there. It also exposes you to radiation. You should talk to your doctor about the benefits and drawbacks of mammograms. Together, you can decide when to start and how often to have a mammogram.

Now while it’s difficult to figure out physicians by seeing only images of X-rays that weather the tumour is toxic or not, training a machine learning model to identify a tumour can be of great help.

**Dataset Description**

The Projects Features Detection of Breast Cancer Using Machine Learning. It has been tested that while several machine learning models exist, Support Vector Machine or SVM in short is reported to have the highest accuracy of (approximately 97%) in detecting breast cancer.

The dataset used in this project is from the Breast Cancer Wisconsin (Diagnostic) Data Set, however, it can be directly accessed from Scikit learn library's collection of datasets as... sklearn.datasets.load\_breast\_cancer

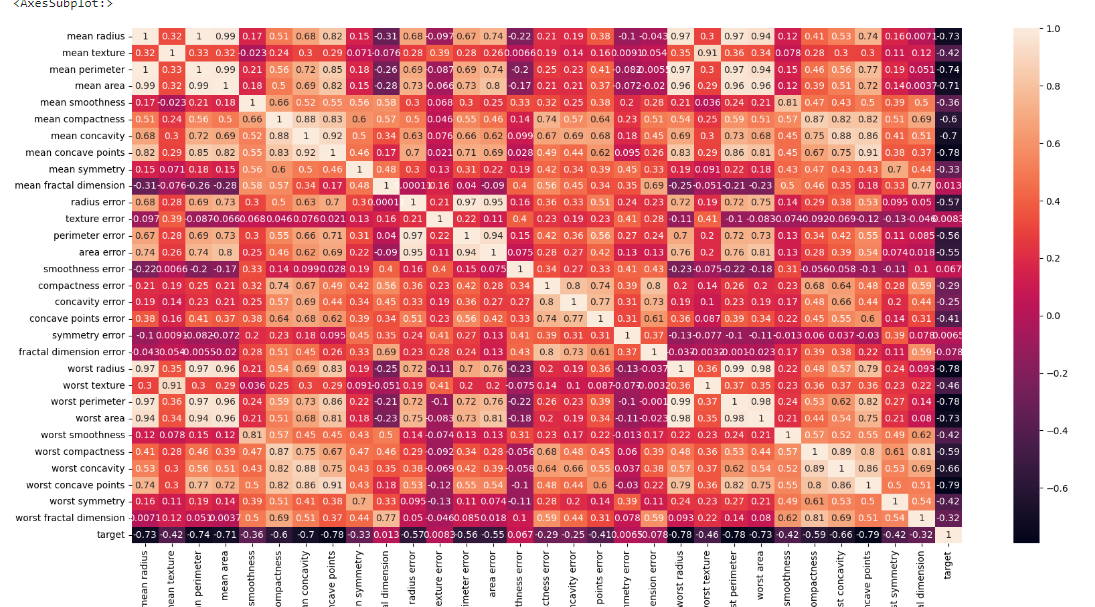
...also, CSV file of data has been externally loaded in the repo :)

**What is SVM?**

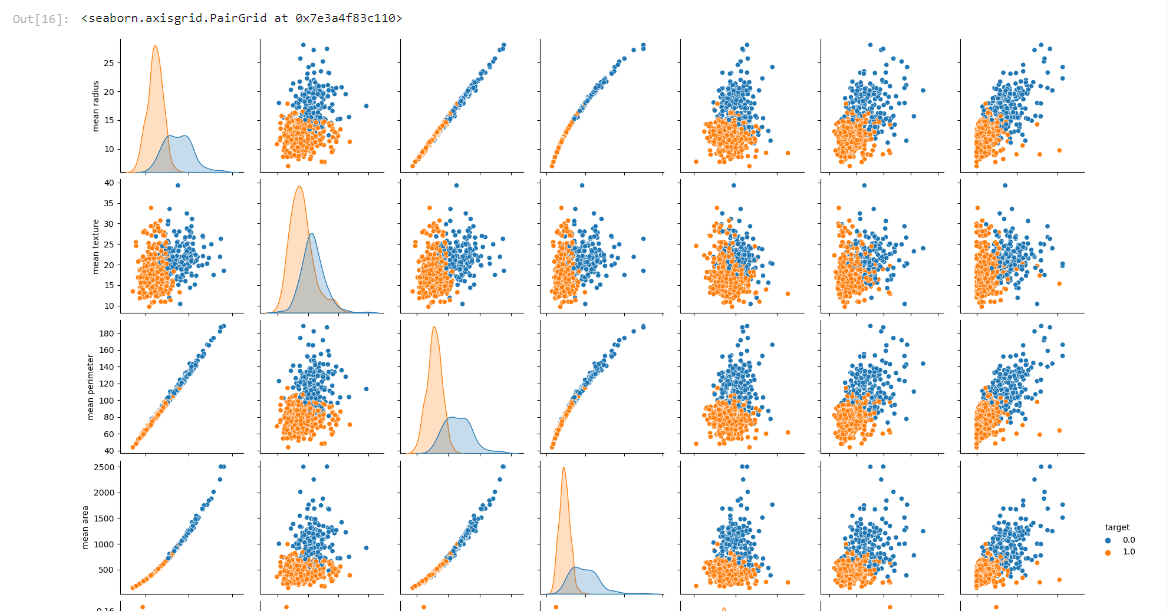
* The Projects Features Detection of Breast Cancer Using Machine Learning. It has been tested that while several machine learning models exist, Support Vector Machine or SVM in short is reported to have the highest accuracy of (approximately 97%) in detecting breast cancer.
* Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.
* The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
* Diagram

  Description automatically generated

**DATA VISUALIZATION**



Correlation matrix of all features of the dataset

A picture containing text

Description automatically generated

Pairplot showing scatter plot, SDE plot relations between 'mean radius', 'mean texture', 'mean perimeter', 'mean area',

'mean smoothness', 'mean compactness' ,'mean concavity' variables with the ‘target’ variable as hue.

**Accuracy Metrics**

**What is Confusion Matrix and why you need it?**

Well, it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.

It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC-ROC curves.

**True Positive:**

Interpretation: You predicted positive and it’s true.

You predicted that a woman is pregnant and she actually is.

**True Negative:**

Interpretation: You predicted negative and it’s true.

You predicted that a man is not pregnant and he actually is not.

**False Positive: (Type 1 Error)**

Interpretation: You predicted positive and it’s false.

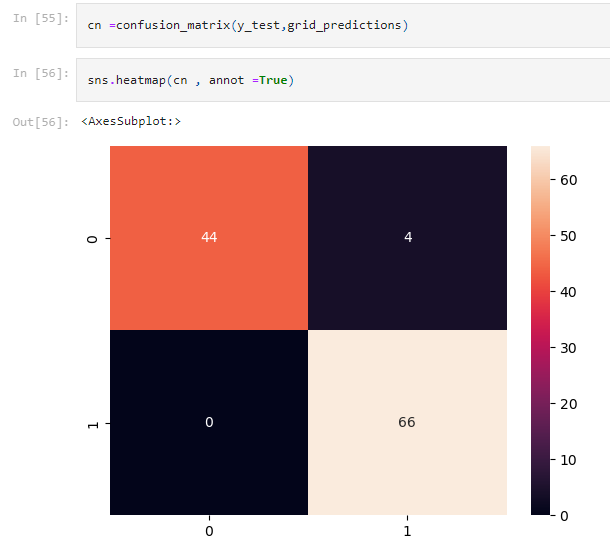
You predicted that a man is pregnant but he actually is not.

**False Negative: (Type 2 Error)**

Interpretation: You predicted negative and it’s false.

Table

Description automatically generated



Confusion matrix screenshot from the project

**RESULTS**

An accuracy of 96% was achieved by using the SVM model and after the normalization technique after optimisation of C and Gamma parameters, it was increased to a value of 97%.

precision recall f1-score support

0.0 1.00 0.92 0.96 48

1.0 0.94 1.00 0.97 66

accuracy 0.96 114

macro avg 0.97 0.96 0.96 114

weighted avg 0.97 0.96 0.96 114

**How to access this project.**

Fork the Repository and clone it in ur PC, voila its urs now use it your own way i hope u will do even cooler things ;)