

# PROJECT REPORT

PROJECT NAME: Breast Cancer Detection and Prediction using Machine Learning

GUIDENCE: Letsupgrade

PERFORM BY

Prasad Balasaheb Nehe

## ABSTRACT

According to the world health organization (WHO) Breast cancer is the most frequent cancer among women, impacting 2.1 million women each year, and also causes the greatest number of cancer- related deaths among women. In 2018, it is estimated that 627,000 women died from breast cancer – that is approximately 15% of all cancer deaths among women. While breast cancer rates are higher among women in more developed regions, rates are increasing in nearly every region globally. In order to improve breast cancer outcomes and survival, early detection is critical. There are two early detection strategies for breast cancer: early diagnosis and screening. Limited resource settings with weak health systems where the majority of women are diagnosed in late stages should prioritize early diagnosis programs based on awareness of early signs and symptoms and prompt referral to diagnosis and treatment. Early diagnosis strategies focus on providing timely access to cancer treatment by reducing barriers to care and/or improving access to effective diagnosis services. The goal is to increase the proportion of breast cancers identified at an early stage, allowing for more effective treatment to be used and reducing the risks of death from breast cancer. Since early detection of cancer is key to effective treatment of breast

cancer, we use various machine learning algorithms to predict if a tumor is benign or malignant, based on the features provided by the data. Cancer, we use various machine learning algorithms to predict if a tumor is benign or malignant, based on the features provided by the data.

# INTRODUCTION

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society.

The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical

features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modeling.

## **Some Risk Factors for Breast Cancer**

The following are some of the known risk factors for breast cancer. However, most cases of breast cancer cannot be linked to a specific cause. Talk to your doctor about your specific risk.

**Age.** The chance of getting breast cancer increases as women age. Nearly 80 percent of breast cancers are found in women over the age of 50.

**Personal history of breast cancer.** A woman who has had breast cancer in one breast is at an increased risk of developing cancer in her other breast.

**Family history of breast cancer.** A woman has a higher risk of breast cancer if her mother, sister or daughter had breast cancer, especially at a young age (before 40). Having other relatives with breast cancer may also raise the risk.

**Genetic factors.** Women with certain genetic mutations, including changes to the BRCA1 and BRCA2 genes, are at higher risk of developing breast cancer during their lifetime. Other gene changes may raise breast cancer risk as well.

**Childbearing and menstrual history.** The older a woman is when she has her first child, the greater her risk of breast cancer. Also, at higher risk are:

1. Women who menstruate for the first time at an early age (before 12)
2. Women who go through menopause late (after age 55)
3. Women who've never had children

## **Role of Machine Learning In 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 its difficult to figure out for physicians by seeing only images of x-ray that weather the tumor is toxic or not training a machine learning model according to the identification of tumor can be of great help.

## LITERATURE SURVEY

Twenty-four recent research articles have been reviewed to explore the computational methods to predict breast cancer. The summaries of them are presented below. Chourasia et al. developed prediction models of benign and malignant breast cancer. Wisconsin breast cancer data set was used. The dataset contained 699 instances, two classes (malignant and benign), and nine integer-valued clinical attributes such as uniformity of cell size. The researchers removed the 16 instances with missing values from the data set to become the data set of 683 instances. The benign were 458 (65.5%) and malignant were 241 (34.5%). The experiment was analyzed by the Waikato Environment for Knowledge Analysis (WEKA). Naive Bayes, RBF Network, and J48 are the three most popular data mining algorithms were used to develop the prediction models. The researchers used 10-fold cross-validation methods to measure the unbiased

## PROBLEM STATEMENT & DISCUSSION

Breast Cancer is one of the leading cancers developed in many countries including India. Though the endurance rate is high – with early diagnosis 97% women can survive for more than 5 years. Statistically, the death toll due to this disease has increased drastically in last few decades. The main issue pertaining to its cure is early recognition. Hence, apart from medicinal solutions some Data Science solution needs to be integrated for resolving the death causing issue.

This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this I have used machine learning classification methods to fit a function that can predict the discrete class of new input

## PROPOSED SOLUTION & RESULT ANALYSIS

In this project we will use Data Mining and Machine Learning Algorithms to detect breast cancer, based off of data. Breast Cancer (BC) is a common cancer for women around the world.

Early detection of BC can greatly improve prognosis and survival chances by promoting clinical treatment to patients.

We will use the UCI Machine Learning Repository for breast cancer dataset.

Url: [http://archive.ics.uci.edu/ml/datasets/breast+cancer+ Wisconsin+%28diagnostic%29](http://archive.ics.uci.edu/ml/datasets/breast+cancer+Wisconsin+%28diagnostic%29)

The dataset used in this story is publicly available and was created by Dr. William H. Walberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. To create the dataset Dr. Walberg used fluid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Cyst, which is capable of perform the analysis of cytological features based on a digital scan. The program uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, then it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector Attribute Information:

1. ID number 2) Diagnosis (M = malignant, B = benign) 3–32)

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

2. radius (mean of distances from center to points on the perimeter)
3. texture (standard deviation of gray-scale values)
4. perimeter
5. area
6. smoothness (local variation in radius lengths)
7. compactness ( $\text{perimeter}^2 / \text{area} - 1.0$ )
8. concavity (severity of concave portions of the contour)
9. concave points (number of concave portions of the contour)

10. symmetry

11. fractal dimension (“coastline approximation” — 1)

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.

## DATA MINING AND MACHINE LEARNING

The term “data mining” is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence (e.g., machine learning) and business intelligence. The book Data mining: Practical machine learning tools and techniques with Java (which covers mostly machine learning material) was originally to be named just Practical machine learning, and the term data mining was only added for marketing reasons. Often the more general terms (large scale) data analysis and analytics – or, when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

In this project we use the following machine learning algorithms:

**Decision tree algorithms:** Decision tree algorithms are successful machine learning classification techniques. They are the supervised learning methods which use information gained and pruned to improve results. Moreover, decision tree algorithms are commonly used for classification in many researches, for example, in the medicine area and health issues. There are many kinds of decision tree algorithms such as ID3 and C4.5. However, J48 is the most popular decision tree algorithm. J48 is the implementation of an improved version of C4.5 and is an extension of ID3.

**Support Vector Machine (SVM):** It is a supervised learning method derived from statistical learning theory for the classification of both linear and nonlinear data. SVM classifies data into two classes over a hyperplane at the same time avoiding over-fitting the data by maximizing the margin of hyperplane separating

**Logistic regression:** In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one. Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable (target) is categorical.

# Data Exploration

We will first go with importing the necessary libraries and import our dataset to colab.research.google.com.

We can examine the data set using the pandas' head() method.

df.head() {first 5 rows of the data}

Out[7]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	te
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	

5 rows x 33 columns

In [9]: data.columns

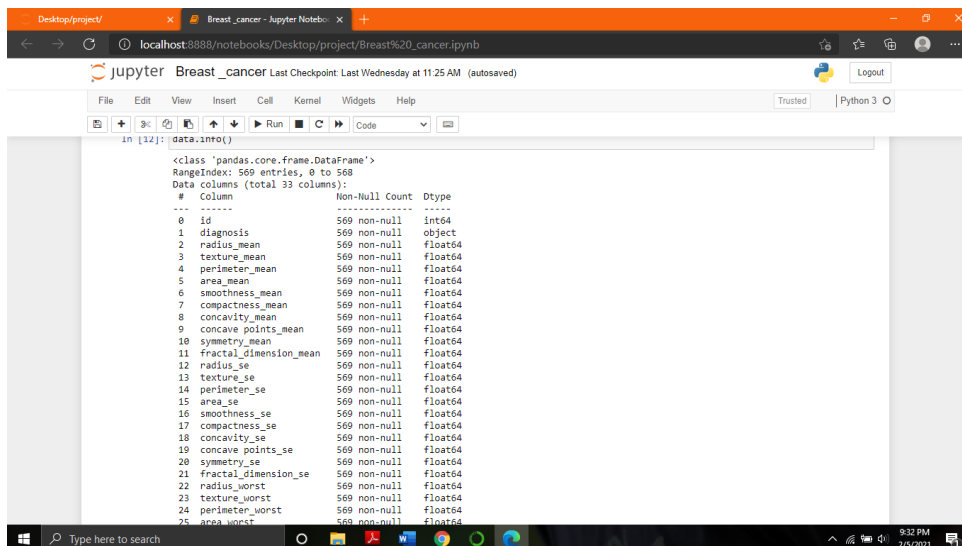
estimate of the three prediction models for performance comparison purposes. The models' performance evaluation was presented based on the methods' effectiveness and accuracy. Experimental results showed that the Naive Bayes had gained the best performance with a classification accuracy of 97.36%; followed by RBF Network with a classification accuracy of 96.77% and the J48 was the third with a classification accuracy of 93.41%. In addition, the researchers conducted sensitivity analysis and specificity analysis of the three algorithms to gain insight into the relative contribution of the independent variables to predict survival. The sensitivity results indicated that the prognosis factor

We can observe that the data set contain 569 rows and 33 columns. 'Diagnosis' is the column which we are going to predict, which says if the cancer is M = malignant or B = benign. 1 means the cancer is malignant and 0 means benign. We can identify that out of the 569 persons, 357 are labeled as B (benign) and 212 as M (malignant).

Each row represents a patient and 33 features on the 569 patients.

The last column Unnamed: 32 has Nan values so we need to remove that column with empty values.

So, we count the number of empty columns and drop the columns with empty values.



```
Desktop/project/ x Breast_cancer - Jupyter Notebo... x +
localhost:8888/notebooks/Desktop/project/Breast%20cancer.ipynb
jupyter Breast_cancer Last Checkpoint: Last Wednesday at 11:25 AM (autosaved) Logout
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
In [12]: data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 33 columns):
# Column Non-Null Count Dtype
---
0 id 569 non-null int64
1 diagnosis 569 non-null object
2 radius_mean 569 non-null float64
3 texture_mean 569 non-null float64
4 perimeter_mean 569 non-null float64
5 area_mean 569 non-null float64
6 smoothness_mean 569 non-null float64
7 compactness_mean 569 non-null float64
8 concavity_mean 569 non-null float64
9 concave points_mean 569 non-null float64
10 symmetry_mean 569 non-null float64
11 fractal_dimension_mean 569 non-null float64
12 radius_se 569 non-null float64
13 texture_se 569 non-null float64
14 perimeter_se 569 non-null float64
15 area_se 569 non-null float64
16 smoothness_se 569 non-null float64
17 compactness_se 569 non-null float64
18 concavity_se 569 non-null float64
19 concave points_se 569 non-null float64
20 symmetry_se 569 non-null float64
21 fractal_dimension_se 569 non-null float64
22 radius_worst 569 non-null float64
23 texture_worst 569 non-null float64
24 perimeter_worst 569 non-null float64
25 area_worst 569 non-null float64
26 Unnamed: 32 0 non-null object
```



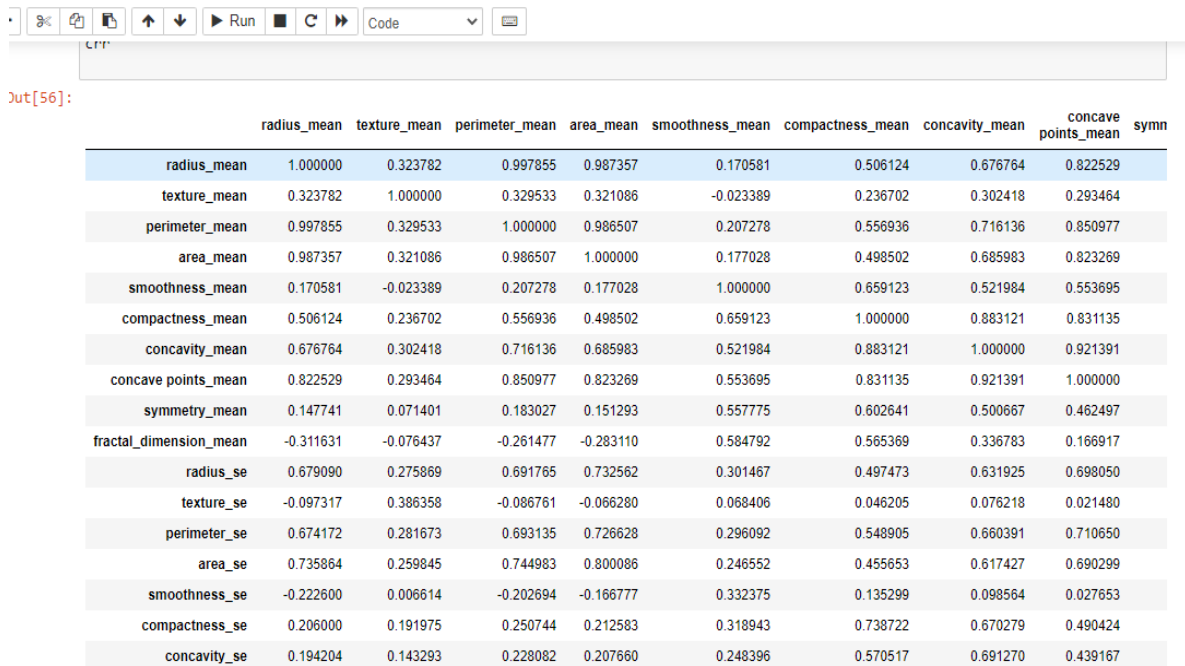
So column Unnamed: 32 has 569 missing values so we drop it.

So the new shape of the data is (569, 32) which means 569 rows and 32 columns.

Now we can see the number of Malignant (M) (harmful) or Benign (B) cells (not harmful) cells and plot it in a graph.

We can see that id column acts as the identifier of the patient and it is of integer type and it cannot be used as a feature to predict the tumor.

Next we encode categorical data values (Transforming categorical data/ Strings to integers)

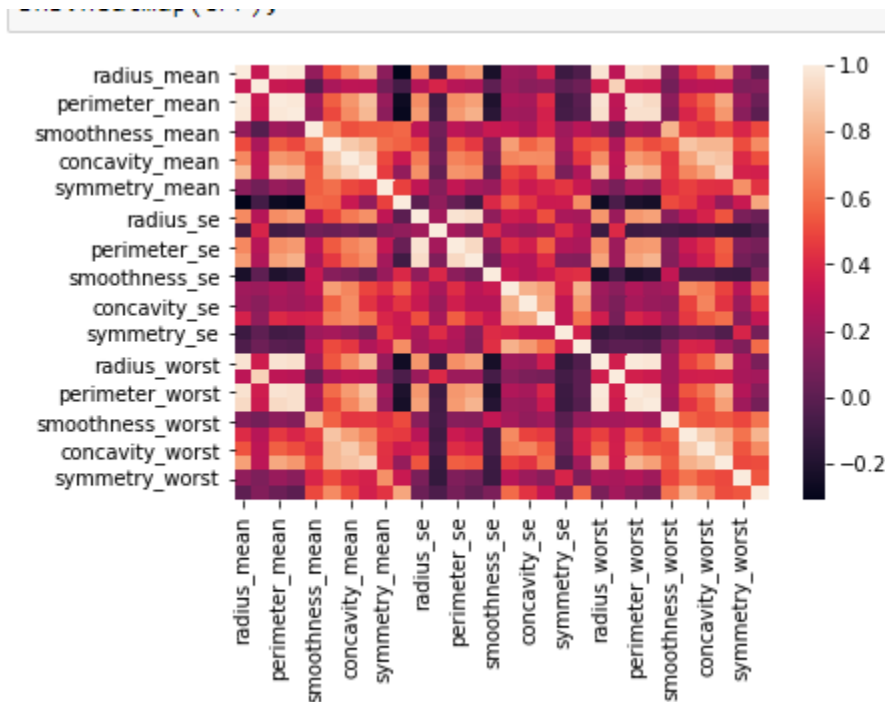


Out[56]:

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symm
radius_mean	1.000000	0.323782	0.997855	0.987357	0.170581	0.506124	0.676764	0.822529	
texture_mean	0.323782	1.000000	0.329533	0.321086	-0.023389	0.236702	0.302418	0.293464	
perimeter_mean	0.997855	0.329533	1.000000	0.986507	0.207278	0.556936	0.716136	0.850977	
area_mean	0.987357	0.321086	0.986507	1.000000	0.177028	0.498502	0.685983	0.823269	
smoothness_mean	0.170581	-0.023389	0.207278	0.177028	1.000000	0.659123	0.521984	0.553695	
compactness_mean	0.506124	0.236702	0.556936	0.498502	0.659123	1.000000	0.883121	0.831135	
concavity_mean	0.676764	0.302418	0.716136	0.685983	0.521984	0.883121	1.000000	0.921391	
concave points_mean	0.822529	0.293464	0.850977	0.823269	0.553695	0.831135	0.921391	1.000000	
symmetry_mean	0.147741	0.071401	0.183027	0.151293	0.557775	0.602641	0.500667	0.462497	
fractal_dimension_mean	-0.311631	-0.076437	-0.261477	-0.283110	0.584792	0.565369	0.336783	0.166917	
radius_se	0.679090	0.275869	0.691765	0.732562	0.301467	0.497473	0.631925	0.698050	
texture_se	-0.097317	0.386358	-0.086761	-0.066280	0.068406	0.046205	0.076218	0.021480	
perimeter_se	0.674172	0.281673	0.693135	0.726628	0.296092	0.548905	0.660391	0.710650	
area_se	0.735864	0.259845	0.744983	0.800086	0.246552	0.455653	0.617427	0.690299	
smoothness_se	-0.222600	0.006614	-0.202694	-0.166777	0.332375	0.135299	0.098564	0.027653	
compactness_se	0.206000	0.191975	0.250744	0.212583	0.318943	0.738722	0.670279	0.490424	
concavity_se	0.194204	0.143293	0.228082	0.207660	0.248396	0.570517	0.691270	0.439167	

Here the value 1 represents Malignant (M) (harmful) and value 0 represents Benign (B) cells (not harmful) cells.

Now we visualize a correlation between the different attributes.



In this heat map we can see how much one column influences all the other columns(e.g radius mean has 32% influence on texture mean).**Training and Testing**

Next we split the datasets into independent (X) and dependent (Y) datasets. `X = df.iloc[:, 2:31].values`

`Y = df.iloc[:, 1].values`

They are of type array.

The dependent data set(Y) has the diagnosis whether the patient has cancer and the independent dataset(X) has the features that are used to predict the outcome.

Now we split the dataset into 75% training and 25%testing and use different machine learning models such as logistic regression , random forest classifier, decision tree to the training set.

```
In [75]: y_test.shape
Out[75]: (171,)

In [76]: from sklearn.preprocessing import StandardScaler
ss = StandardScaler()
X_train = ss.fit_transform(X_train)
X_test = ss.transform(X_test)

In [77]: X_train
Out[77]: array([[ 1.83083287,  0.43037905,  1.87998089, ...,  1.43208963,
                -0.71185545,  0.30894129],
                [-0.82726482,  2.34731259, -0.87502275, ..., -1.75416862,
                -2.21050456, -1.40375336],
                [ 0.35316834,  0.78877253,  0.33688337, ...,  0.47153757,
                -0.5872413 , -0.89282344],
```

## Prediction of Model

```
pred = model[5].predict(X_test) print(pred)
```

So here we printed the predictions. The first data shows the actual result of which patient had cancer and the second data is the one predicted by the model.

The accuracy of the model is 96.5% so we can see a few wrong predictions but mostly this model is successful in predicting a tumor Malignant (M) (harmful) or Benign (B) (not harmful) based upon the features provided in the data and the training given.

```
Out[91]: SVC()
```

```
In [92]: y_conclude = svc.predict(X_test)
y_conclude
```

```
Out[92]: array([0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1,
                0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1,
                1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0,
                0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0,
                0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0,
                0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0,
                1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0], dtype=int64)
```

Accuracy of different model:

```
Out[118]:
```

	Algorithm	Accuracy
0	decision tree classifier	0.947368
0	logistic Regression Method	0.982456
0	random forest Classifier Method	0.970760
0	Support Vector Classifier Method	0.982456

## **CONCLUSION & FUTURE SCOPE**

In this project in python, we learned to build a breast cancer tumor predictor on the Wisconsin dataset and created graphs and results for the same. It has been observed that a good dataset provides better accuracy. Selection of appropriate algorithms with good home dataset will lead to the development of prediction systems. These systems can assist in proper treatment methods for a patient diagnosed with breast cancer. There are many treatments for a patient based on breast cancer stage; data mining and machine learning can be a very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable databases.

## REFERENCES

1. <https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/>
2. <https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3>
3. Original data Set: <http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29>
4. Confusion Matrix:  
<https://tatwan.github.io/How-To-Plot-A-Confusion-Matrix-In-Python/>
5. [https://seaborn.pydata.org/tutorial/axis\\_grids.html](https://seaborn.pydata.org/tutorial/axis_grids.html)
6. <https://seaborn.pydata.org/generated/seaborn.pairplot.html>
7. <https://seaborn.pydata.org/generated/seaborn.heatmap.html>