A PROJECT REPORT ON

PREDICTION OF BREAST CANCER

BY USING

MACHINE LEARNING

SUBMITTED BY:

VISHWAS ANAND
ANUBHAB SAHU
SANGEETA BARMAN
NIKITA PANDA

PROJECT GUIDE:

SOFIKUL MULLICK

Submission Date: 02 July 2019

CONTENTS

Acknowledgement

The achievement that is associated with the successful completion of any task would be incomplete without mentioning the names of those people whose endless cooperation made it possible. We take this opportunity to express our deep gratitude towards our project mentor, **Mr. Sofikul Mullick** for giving such valuable suggestions, guidance and encouragement during the development of this project work.

ABSTRACT

The discovery of knowledge from medical datasets is important in order to make effective medical diagnosis. With the emerging increase of breast cancer , that recently affects around 2 million people every year, of which more than one-third go undetected in early stage, a strong need for supporting the medical decision- making process is generated.

Breast cancer is a major public health challenge worldwide. Typically, the cancer forms in either the lobules or ducts of the breast. There are five main stages of breast cancer.

Logistic regression and SVM regression methods are now popularly used for prediction in Machine Learning.using both the logistic and SVM regression method to predict the result and then opting for the best accuracy one result in better and precise prediction.

INTRODUCTION

a. CANCER:-Cancer is a broad term. It describes the disease that results when cellular changes cause the uncontrolled growth and division of cells. Some types of cancer cause rapid cell growth, while others cause cells to grow and divide at a slower rate.

Certain forms of cancer result in visible growths called tumors, while others, such as leukemia, do not.

Most of the body's cells have specific functions and fixed lifespans. While it may sound like a bad thing, cell death is part of a natural and beneficial phenomenon called apoptosis.

A cell receives instructions to die so that the body can replace it with a newer cell that functions better. Cancerous cells lack the components that instruct them to stop dividing and to die.

As a result, they build up in the body, using oxygen and nutrients that would usually nourish other cells. Cancerous cells can form tumors, impair the immune system and cause other changes that prevent the body from functioning regularly.

Cancerous cells may appear in one area, then spread via the lymph nodes. These are clusters of immune cells located throughout the body.

i. Types of Cancer:-The main types of diabetes are described below:

- bladder

- colon and rectal
- endometrial
- kidney
- leukemia
- liver
- melanoma
- non-Hodgkin's lymphoma
- pancreatic
- thyroid
- breast

ii. BREAST CANCER:-

Breast cancer is cancer that forms in the cells of the breasts.

After skin cancer, breast cancer is the most common cancer diagnosed in women in the United States. Breast cancer can occur in both men and women, but it's far more common in women.

iii. SYMPTOMS:-

- A breast lump or thickening that feels different from the surrounding tissue
- Change in the size, shape or appearance of a breast
- Changes to the skin over the breast, such as dimpling
- A newly inverted nipple
- Peeling, scaling, crusting or flaking of the pigmented area of skin surrounding the nipple (areola) or breast skin

b. MACHINE LEARNING:-

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed., it provides set of methods that can detect patterns in the data and use the patterns to generate future predictions.

Because of new computing technologies, machine learning today is not like machine learning of the past. With the rise of Machine Learning approaches we have the ability to find a solution to this issue, we have developed a system using data mining which has the ability to predict whether the patient has diabetes or not.

Furthermore, predicting the disease early leads to treating the patients before it becomes critical.The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results.

There are three core types of machine learning- supervised learning, unsupervised learning, and reinforcement learning.

iv. Supervised Learning:-The main goal in supervised learning is to learn a model from labeled training data that allows us to make predictions about unseen or future data. Here, the term **supervised** refers to a set of samples where the desired output signals (labels) are already known.

v. Semi-supervised Learning:-It uses both labeled and unlabeled data for training-typically a small amount of labeled data and the large amount of unlabeled data(because unlabeled data is less expensive and take less effort to acquire).

vi. Reinforcement Learning:-This learning used for robotics,gaming and navigation.With reinforcement learning,the algorithm discovers through

trial and error which actions yield the greatest rewards

TYPES OF ALGORITHMS:-

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. SVM
5. Naive Bayes
6. kNN
7. K-Means
8. Random Forest
9. Dimensionality Reduction Algorithms
10. Gradient Boosting algorithms

PROJECT OBJECTIVE:-

- The main objective of this project is to predict the breast cancer.

- The work focuses on to predict cancer using logistic model of Regression.

- This integrated technique of classification gives a promising classification results with utmost accuracy rate.

- For detecting a disease a number of test should be required from the patient.

- Cells are the basic structural unit in the body.

- cells are responsible for the growth of a person.

- However if the cells see a rapid growth or slow growth,it can result in tumor.This disease is referred to as cancer.

- In early the ability to diagnose cancer plays an important role for the patient's treatment process.

Problem Statement and Description :

Prediction of breast cancer using Logistic Regression : To identify whether a given person in dataset will be suffering from breast cancer or not, it will be done on the basis of attribute values.

Dataset contains all the details of person 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 and other details of the cell of a person.

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 and others exceeding a specific value may contribute to identify whether a person is diabetic, non diabetic or prediabetic.

The aim of prediction of breast cancer is to make aware people about breast cancer and what it takes to treat it and gives the power to control. It makes necessary chances to improve lifestyle. The proposed Logistic Regression will predict the person having breast cancer or not.

The Logistic Regression Algorithm

Logistic Regression is one of the most used Machine Learning algorithms for binary classification. It is a simple Algorithm that you can use as a performance baseline, it is easy to implement and it will do well enough in many tasks. Therefore every Machine Learning engineer should be familiar with its concepts. The building block concepts of Logistic Regression can also be helpful in deep learning while building neural networks.

Like many other machine learning techniques, it is borrowed from the field of statistics and despite its name, it is not an algorithm for regression problems, where you want to predict a continuous outcome. Instead, Logistic Regression is the go-to method for binary classification. It gives you a discrete binary outcome between 0 and 1. To say it in simpler words, it's outcome is either one thing or another.

A simple example of a Logistic Regression problem would be an algorithm used for diabetes detection that takes an input and should tell if a patient has diabetes (1) or not (0).

How it works

Logistic Regression measures the relationship between the dependent variable (our label, what we want to predict) and one or more independent variables (our features), by estimating probabilities using it's underlying logistic function.

These probabilities must then be transformed into binary values in order to actually make a prediction.

Advantages of logistic regression:

- it is more robust: the independent variables don't have to be normally distributed, or have equal variance in each group.
- It does not assume a linear relationship between the IV and DV.
- It may handle nonlinear effects.
- You can add explicit interaction and power terms.
- The DV need not be normally distributed.
- There is no homogeneity of variance assumption.
- Normally distributed error terms are not assumed.
- It does not require that the independents be interval.
- It does not require that the independents be unbounded.

Disadvantages of logistic regression:

- Identifying Independent Variables
- Limited Outcome Variables
- Independent Observations Required
- Overfitting the Model

SVM:-

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

KERNEL TUNING PARAMETER:-

The learning of the hyperplane in linear SVM is done by transforming the problem using some linear algebra. This is where the kernel plays role.

For linear kernel the equation for prediction for a new input using the dot product between the input (x) and each support vector (xi) is calculated as follows:

f(x) = B(0) + sum(ai * (x,xi))

This is an equation that involves calculating the inner products of a new input vector (x) with all support vectors in training data. The coefficients B0 and ai (for each input) must be estimated from the training data by the learning algorithm.

The polynomial kernel can be written as $K(x,xi) = 1 + sum(x * xi)^d$ and exponential as $K(x,xi) = exp(-gamma * sum((x — xi^2))$.

Project:

#importing the essential libraries

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
from sklearn.datasets import load_breast_cancercancer =
load_breast_cancer()
```

```
cancer.keys()
```

```
dict_keys(['data', 'target', 'target_names', 'DESCR', 'feature_names', 'filename'])
```

```
In [8]: print(cancer['target'])
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 1 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 1 0 0 1 1 1 1 0 1 0 0 1 1 1 1 0 1 0 0
 1 0 1 0 0 1 1 1 0 0 1 0 0 0 1 1 1 0 1 1 0 0 1 1 1 0 0 1 1 1 1 0 1 1 0 1 1
 1 1 1 1 1 1 0 0 0 1 0 0 1 1 1 0 0 1 0 1 0 0 1 0 0 1 1 0 1 1 0 1 1 1 1 0 1
 1 1 1 1 1 1 1 1 0 1 1 1 1 0 0 1 0 1 1 0 0 1 1 0 0 1 1 1 1 0 1 1 0 0 0 1 0
 1 0 1 1 1 0 1 1 0 0 1 0 0 0 0 1 0 0 0 1 0 1 0 1 1 0 1 0 0 0 0 1 1 0 0 1 1
 1 0 1 1 1 1 1 1 0 0 1 1 0 1 1 0 0 1 0 1 1 1 1 0 1 1 1 1 1 0 1 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 1 0 1 1 0 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1
 1 0 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 0 1 1 1 1 0 0 0 1 1
 1 1 0 1 0 1 0 1 1 1 0 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0
 0 1 0 0 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 0 1 1 0 0 1 1 1 1 1 1 0 1 1 1 1 1 1
 1 0 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 0 1 1 1 1 1 0 1 1
 0 1 0 1 1 0 1 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1
 1 1 1 1 1 1 0 1 0 1 1 0 1 1 1 1 1 1 0 0 1 0 1 0 1 1 1 1 1 1 0 1 1 0 1 0 1 0 0
 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 0 0 0 0 0 0 1]
```

```
print(cancer['target_names'])
      ['malignant' 'benign']
print(cancer['feature_names'])
        ['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']

cancer['data'].shape
           (569, 30)
df_cancer = pd.DataFrame(np.c_[cancer['data'], cancer['target']], columns= np.ap
```

In [11]:
```
df_cancer.head()
```

Out[11]:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	di
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	

5 rows × 31 columns

In [12]:
```
df_cancer.tail()
```

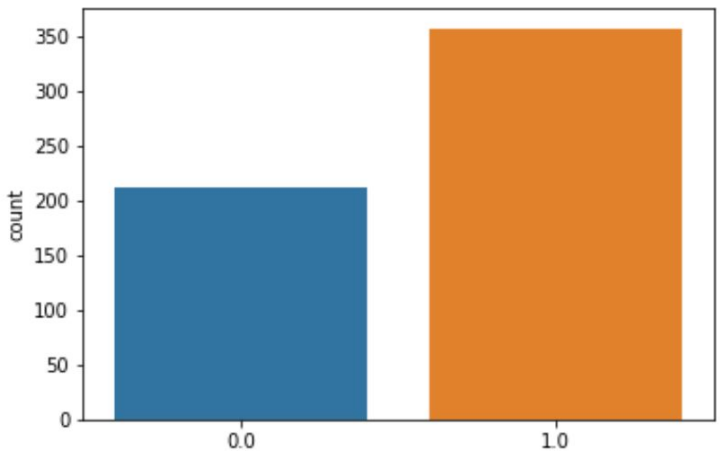
Out[12]:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587

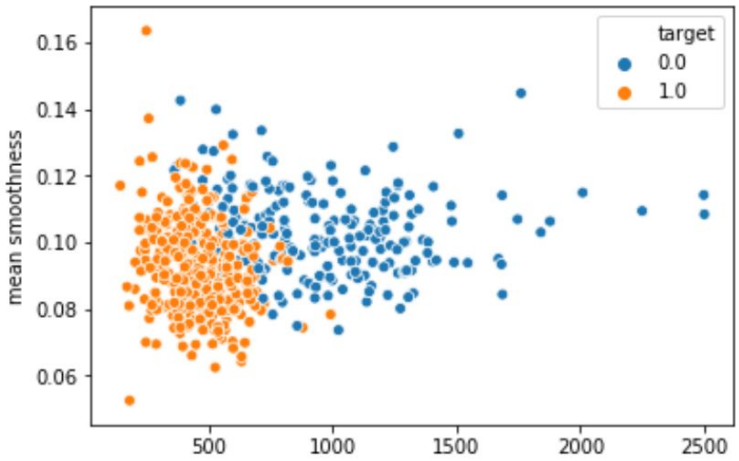
5 rows × 31 columns

In [17]:
```
sns.countplot(df_cancer['target'])
```

Out[17]: `<matplotlib.axes._subplots.AxesSubplot at 0x1e5a6358780>`



```
In [18]: sns.scatterplot(x='mean area',y='mean smoothness',hue='target',data =df_cancer)
Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x1e5a6358ba8>
```




```
In [22]: y= df_cancer['target']

In [23]: y
Out[23]: 0      0.0
         1      0.0
         2      0.0
         3      0.0
         4      0.0
         5      0.0
         6      0.0
         7      0.0
         8      0.0
         9      0.0
         10     0.0
         11     0.0
         12     0.0
         13     0.0
```

from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=5)

In [27]:
```
y_train
```

Out[27]:
```
306    1.0
410    1.0
197    0.0
376    1.0
244    0.0
299    1.0
312    1.0
331    1.0
317    0.0
341    1.0
156    0.0
71     1.0
218    0.0
344    1.0
247    1.0
212    0.0
```

In [28]:
```
x_test
```

Out[28]:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry
28	15.300	25.27	102.40	732.4	0.10820	0.16970	0.168300	0.087510	0.1926
163	12.340	22.22	79.85	464.5	0.10120	0.10150	0.053700	0.028220	0.1551
123	14.500	10.89	94.28	640.7	0.11010	0.10990	0.088420	0.057780	0.1856
361	13.300	21.57	85.24	546.1	0.08582	0.06373	0.033440	0.024240	0.1815
549	10.820	24.21	68.89	361.6	0.08192	0.06602	0.015480	0.008160	0.1976
339	23.510	24.27	155.10	1747.0	0.10690	0.12830	0.230800	0.141000	0.1797
286	11.940	20.76	77.87	441.0	0.08605	0.10110	0.065740	0.037910	0.1588
354	11.140	14.07	71.24	384.6	0.07274	0.06064	0.045050	0.014710	0.1690
421	14.690	13.98	98.22	656.1	0.10310	0.18360	0.145000	0.063000	0.2086

In [29]:
```
y_test
```

Out[29]:
```
28     0.0
163    1.0
123    1.0
361    1.0
549    1.0
339    0.0
286    1.0
354    1.0
421    1.0
124    1.0
543    1.0
537    1.0
567    0.0
555    1.0
511    1.0
```

TRAINING THE MODEL USING SVM

In [30]:
```
from sklearn.svm import SVC
```

In [31]:
```
from sklearn.metrics import classification_report , confusion_matrix
```

In [32]:
```
svc_model= SVC()
```

In [33]:
```
svc_model.fit(x_train,y_train)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\svm\base.py:196: FutureWarni
ng: The default value of gamma will change from 'auto' to 'scale' in version 0.
22 to account better for unscaled features. Set gamma explicitly to 'auto' or
'scale' to avoid this warning.
  "avoid this warning.", FutureWarning)
```

Out[33]:
```
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
  kernel='rbf', max_iter=-1, probability=False, random_state=None,
  shrinking=True, tol=0.001, verbose=False)
```

EVALUATING THE MODEL

In [34]:
```
y_predict =svc_model.predict(x_test)
```

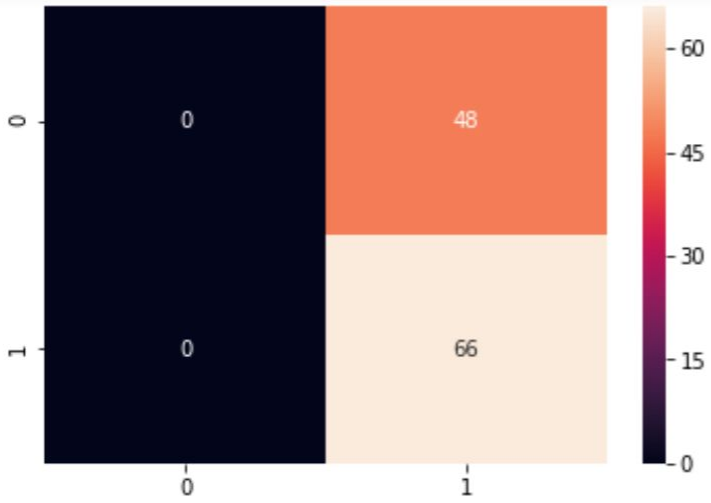
In [35]:
```
y_predict
```

Out[35]:
```
array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.,
       1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.,
       1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.,
       1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.,
       1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.,
       1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.,
       1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.])
```

In [36]:
```
cm = confusion_matrix(y_test,y_predict)
```

In [37]:
```
sns.heatmap(cm ,annot=True)
```

Out[37]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1e5a6ebf0b8>
```



Model Improvisation

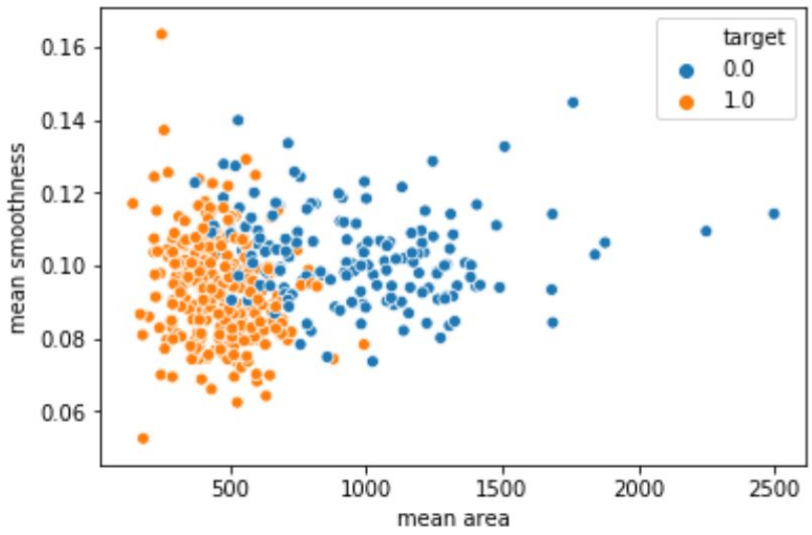
In [38]:
```
min_train =x_train.min()
```

In [39]:
```
range_train =(x_train - min_train).max()
```

In [40]:
```
x_train_scaled =(x_train-min_train)/range_train
```

In [41]:
```
sns.scatterplot(x = x_train['mean area'], y= x_train['mean smoothness'],hue =y_t
```

Out[41]: <matplotlib.axes._subplots.AxesSubplot at 0x1e5a66c9668>



```
In [58]: from sklearn.metrics import accuracy_score

In [59]: accuracy_score(y_test, y_predict)
Out[59]: 0.956140350877193
```

An accuracy of 96% has been achieved after appying the technique of Normalization for Improvisation

```
In [49]: param_grid ={'C':[0.1,1,10,100],'gamma':[1,0.1,0.01,0.001],'kernel':['rbf']}

In [50]: from sklearn.model_selection import GridSearchCV

In [51]: grid=GridSearchCV(SVC(),param_grid,refit=True,verbose=4)

In [53]: grid.best_params_
Out[53]: {'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}

In [54]: grid_predictions=grid.predict(x_test_scaled)

In [55]: cn =confusion_matrix(y_test,grid_predictions)

In [56]: sns.heatmap(cn , annot =True)
Out[56]: <matplotlib.axes._subplots.AxesSubplot at 0x1e5a68dd278>
```


```
In [57]: print(classification_report(y_test,grid_predictions))
```

	precision	recall	f1-score	support
0.0	1.00	0.94	0.97	48
1.0	0.96	1.00	0.98	66
micro avg	0.97	0.97	0.97	114
macro avg	0.98	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

```
In [60]: from sklearn.metrics import accuracy_score
```

Using Logistic Regression

In [63]:
```
X=df_cancer.drop('target', axis=1)
y=df_cancer['target']
```

In [64]:
```
from sklearn.model_selection import train_test_split
```

In [65]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

In [66]:
```
from sklearn.linear_model import LogisticRegression
```

In [67]:
```
logmodel=LogisticRegression()
```

In [69]:
```
predictions=logmodel.predict(X_test)
```

In [70]:
```
print(predictions)
```

```
[1. 0. 0. 1. 0. 1. 1. 0. 0. 1. 1. 1. 1. 1. 0. 0. 1. 1. 1. 1. 1. 1. 0. 0.
 1. 1. 1. 1. 0. 0. 0. 0. 0. 1. 1. 1. 1. 0. 0. 1. 1. 1. 1. 1. 0. 1. 1. 0.
 0. 0. 1. 0. 0. 0. 1. 1. 0. 0. 1. 1. 1. 1. 0. 1. 1. 0. 0. 1. 1. 1. 1. 1.
 1. 1. 0. 1. 1. 1. 1. 0. 1. 1. 1. 1. 1. 0. 0. 0. 1. 1. 1. 1. 1. 1. 1. 0.
 1. 1. 1. 0. 1. 1. 1. 1. 0. 1. 1. 0. 1. 0. 0. 0. 0. 1. 1. 0. 0. 1. 0. 0.
 0. 1. 0. 1. 0. 1. 1. 1. 1. 0. 1. 0. 1. 0. 0. 1. 1. 0. 1. 1. 1. 0. 0. 1.
 0. 1. 1. 0. 1. 1. 1. 0. 1. 1. 1. 0. 1. 0. 0. 1. 1. 1. 1. 1. 0. 1. 1. 1.
 1. 1. 1.]
```

In [71]:
```
from sklearn.metrics import classification_report
```

In [72]:
```
classification_report(y_test, predictions)
```

In [74]:
```
from sklearn.metrics import confusion_matrix
```

In [75]:
```
confusion_matrix(y_test, predictions)
```

Out[75]:
```
array([[ 55,   4],
       [  7, 105]], dtype=int64)
```

In [91]:
```
from sklearn.metrics import accuracy_score
```

In [92]:
```
accuracy_score(y_test, predictions)
```

Out[92]:
```
0.935672514619883
```

Procedures:

```
# Preparing the DataSet

#Splitting dataset
# Training The Model
# Testing the Model

# Table for checking accuracy
Plot the new model
```

EXPERIMENT RESULTS

Datasets Description:

The dataset that is taken for this work is collected from "Pima Indians Diabetes Database" obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. It consists of 768 records of patient data. Here 80% of the data is taken for training and remaining 20% is taken for testing.

The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour postload plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care).

Results:

SVM:-

	precision	recall	f1-score	support
0.0	1.00	0.94	0.97	48
1.0	0.96	1.00	0.98	66
micro avg	0.97	0.97	0.97	114
macro avg	0.98	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

accuracy achieved=96%

LOGISTIC REGRESSION:-

' precision recall f1-score support\n\n 0.0 0.89 0.93 0.91 59\n 1.0 0.96 0.94 0.95 112\n\n micro avg 0.94 0.94 0.94 171\n macro avg 0.93 0.93 0.93 171 n weighted avg 0.94 0.94 0.94 171\n'

accuracy achieved=93.5%

CONCLUSION:

Lately, medical machine learning has gained in interest by the scientific and research communities. cancer is considered as the world's fastest-growing disease. Breast cancer is a type of cancer which occurs in breast cells,

Breast cancer is a disease which occurs when the cells start multiplying at a faster rate which results in formation of lumps and tissues in breasts. We proposed a model in predicting breast cancer by applying logistic regression technique.

In this Logistic Regression is proposed to predict the persons whether diabetic or not.Results have been obtained. For future work, more input features can be used.Moreover, we recommend the proposed models to be tested on a larger dataset.

LIMITATIONS:

The logistic regression model has been used for prediction of diabetes. In this model, we have achieved accuracy level of approximate 93.5% , using different model could help in increasing the accuracy level.
If dataset could be more refined, the accuracy level could have increased.

FUTURE SCOPE:

This project has accuracy of 96% and it can be increased by using different approximations in future.

BIBLIOGRAPHY:

https://towardsdatascience.com/the-logistic-regression-algorithm-75fe48e21cfa

https://www.r-bloggers.com/

http://machinelearningmastery.com/].

https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

https://github.com/gscdit/Breast-Cancer-Detection?files=1