# BREAST CANCER PREDICTION

# USING SVM

##### A PROJECT REPORT

###### Submitted by

|  |  |
| --- | --- |
| **DHANUSH P** | **(732721205012)** |
| **RAVICHANDRAN N** | **(732721205043)** |
| **SANDIP MONDAL** | **(732721205047)** |
| **SURYA S** | **(732721205052)** |

***In Partial Fulfillment of the requirements for the Award of the Degree of***

**BACHELOR OF TECHNOLOGY IN**

**INFORMATION TECHNOLOGY**

### SRI SHANMUGHA COLLEGE OF ENGINEERING AND TECHNOLOGY,

### DECLARATION

We affirm that the project work titled **“BREAST CANCER PREDICTION USING SVM”** being submitted in partial fulfillment for the award of **Bachelor of Technology,** degree is the original work carried out by us. It has not formed the part of any other project work submitted for award of any degree or diploma, either in this or any university.

|  |  |
| --- | --- |
| **DHANUSH P** | **(732721205012)** |
| **RAVICHANDRAN N** | **(732721205043)** |
| **SANDIP MONDAL** | **(732721205047)** |
| **SURYA S** | **(732721205052)** |

**(Signature of the candidates)**

**DHANUSH.P**

##### RAVICHANDRAN.N

##### SANDIP MONDAL

##### SURYA.S

**I certify that the declaration made above by the candidates is true.**

###### (Signature of the Guide) DR.K.MUTHUKANNAN,HOD/IT

**ABSTRACT**

Breast cancer is one of the most common and leading causes of cancer among women. Currently, it has become the common health issue, and its incidence has increased recently. Prior identification is the best way to manage breast cancer results. Computer-aided detection or diagnosis (CAD) systems plays a major role in prior identification of breast cancer and can be used for reduction of death rate among women. The main intention of this paper is to make use of the recent advances in the development of CAD systems and related techniques. The mainstay of the project is to predict whether the person is having breast cancer or not. Machine learning is nothing but training the machines to learn and perform by itself without any explicit program or instruction. So here, predicting whether a person is suffering with breast cancer or not is done with the help of the trained data.

# PREAMBLE

### INTRODUCTION:

Breast cancer is considered a multi-factorial disease and the most common cancer in women worldwide with approximately 30% of all female cancers (i.e. 1.5 million women are diagnosed with breast cancer each year , and 500,000 women die from th Over the past 30 years, this disease has increased, while the death rate has decreased. However, the reduction in mortality due to mammography screening is estimated at 20% and improvement in cancer treatment is estimated at 60%.

Actually, Big data has revolutionized the size of data and also creating value from it Big data has made a big change in BI by analyzing large amount of unstructured, heterogeneous, non-standard and incomplete healthcare data. It does not only forecast but also helps in decision making and is increasingly noticed as breakthrough in ongoing advancement with the goal is to improve the quality of patient care and reduces the healthcare cost. Data mining algorithms applied in healthcare industry play a significant role due to their high performance in predicting, diagnosis of the diseases, reducing costs of medicine, making real time decision to save people's lives. The Most common Data mining modeling goals are classification and prediction which uses several algorithms for the prediction of breast cancer. This paper mainly gives a comparison between the performance of five classifiers: Support Vector Machine (SVM)), Random Forest, Logistic Regression, Decision tree and K-Nearest Neighbors which according to research community are among the most influential data mining algorithms and among the top 10 data mining algorithms . Our objective is to predict and diagnosis breast cancer, using machine-learning algorithms, and find out the most effective based on the performance of each classifier in terms of confusion matrix, accuracy, precision and sensitivity.

### PROBLEM STATEMENT AND DEFINITION

**What is Breast Cancer ?**

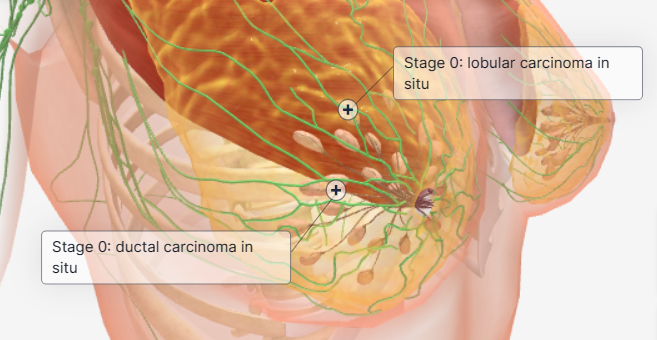
Breast cancer classification considers the tumor’s size, if the cancer is in the lymph nodes, and if it has spread to other organs. The breast cancer classification system also considers the tumor’s grade, genetic makeup, and other characteristics like hormone receptor status.

**Stages in Breast Cancer:**

1. Stage 0
2. Stage 1
3. Stage 2
4. Stage 3
5. Stage 4

**1.Stage 0:**

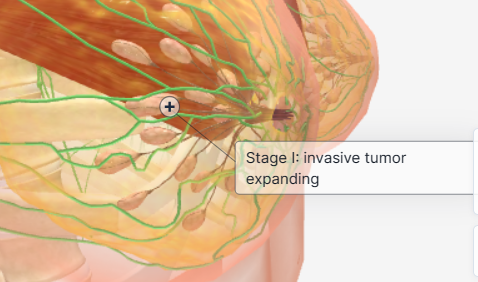
[Stage 0 breast cancer](https://www.verywellhealth.com/breast-cancer-staging-stage-zero-429887) is also called ductal carcinoma in situ (DCIS) or Paget's disease of the breast. It is a precancerous, or noninvasive, abnormal growth in the breast.



Stage 0 breast cancer generally has no signs or symptoms. Often there isn't a lump that you or your healthcare provider can feel. One potential symptom is slight discharge from the nipple. Most of the time, symptoms aren't what lead to a stage 0 breast cancer diagnosis, but instead it is found with a mammogram and breast biopsy.

**Stage 1:**

[Stage 1 (or stage I) breast cancer](https://www.verywellhealth.com/stage-1-breast-cancer-429888) typically has started to grow into surrounding breast tissues. Breast cancers that are stage 1 either have not spread into the lymph nodes (stage 1A) or only spread minimally into the lymph nodes (stage 1B).

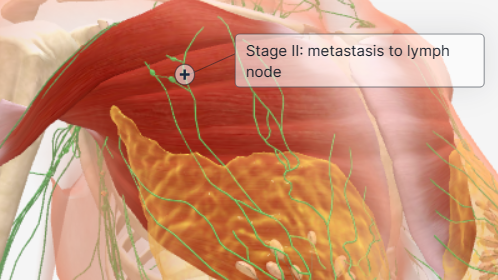


The most common symptom of early breast cancer is a new lump or mass. Most breast lumps are due to benign breast conditions and are not cancer. But a painless, hard mass with irregular edges is more likely to be cancer. Breast cancer lumps can be soft, round, tender, or painful.

* ****Localized****: The cancer hasn’t spread.
* ****Regional****: Cancer has infiltrated nearby tissues or lymph nodes.
* ****Distant****: Cancer has metastasized to other body parts.

**Stage 2:**

[Stage 2 (or stage II) breast cancers](https://www.verywellhealth.com/stage-2-breast-cancer-429889) are typically larger or have spread to more lymph nodes.



Stage 2 breast cancers may cause symptoms like:

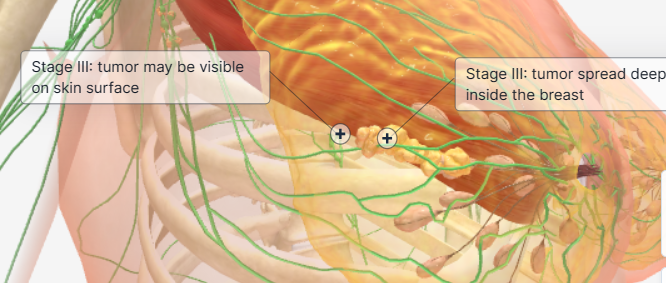
* Breast or nipple pain
* Dry or flaking nipple or breast skin
* Nipple discharge (other than breast milk)

**Stage 3:**

[Stage 3 (or stage III) breast cancer](https://www.verywellhealth.com/stage-3-breast-cancer-429890) is often called invasive breast cancer. It has spread farther than stage 2, involves more lymph nodes, or increased, but has not metastasized beyond the local tissues and organs near the breast.

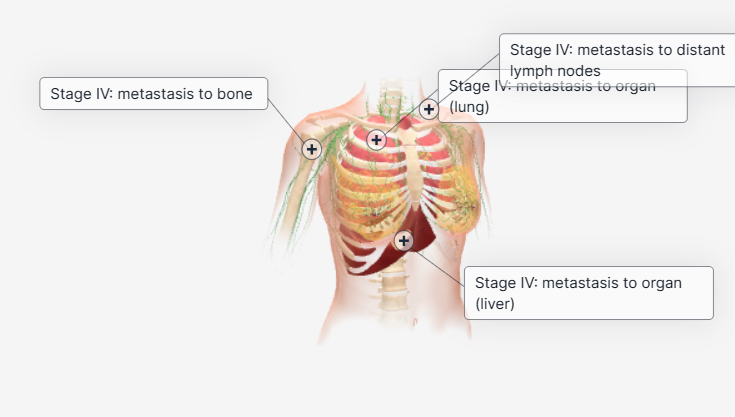
Symptoms of invasive breast cancers are similar to earlier-stage cancers. They can include:

* Breast skin that is red, dimpled, or puckered
* An unexplained rash on the breast
* Fluid coming out of the nipple (that is not breast milk)
* A lump near the breast or armpit
* A change in the shape or feel of the breast



**Stage 4:**

[Stage 4 (or stage IV) breast cancer](https://www.verywellhealth.com/stage-4-breast-cancer-429891) is the deadliest and is not considered curable. It is also called metastatic breast cancer or advanced breast cancer. It has, by definition, spread to organs in other parts of the body. These organs may include the lungs, skin, bones, liver, or brain.



Common symptoms of advanced cancers include:

* Pain
* Fatigue and weakness
* Loss of appetite
* Weight changes
* Nausea and vomiting
* Constipation
* Shortness of breath
* Bone pain that becomes constant, fractures.
* Headaches, nausea, facial numbness, and changes in speech, vision, or balance due to cancer in the brain.
* Pain in the right side of the abdomen, weakness, poor appetite, and yellowing of the eyes and skin from cancer in the liver.
* Lung pain, shortness of breath, and persistent cough.

### PROPOSED SYSTEM

Our proposed system applies supervised machine learning algorithms to detect Breast cancer prediction through the image dataset . Furthermore, we employ these algorithms to implement a classifier using machine learning methods. We identify the most important variables that may lead to higher accuracy in Breast cancer prediction .

Our proposed system applies supervised machine learning algorithms to detect the breast cancer prediction using dataset through image.

We overcome the problem by creating a binary classifier and experimenting with various machine learning techniques to see which fits better.

* Random Forest
* Decision Tree
* SVM(Supporting vector machine)

We will identify the most important variables that may lead to higher accuracy in Breast cancer prediction .

We compare the above models and pick the best one. Which will give Higher Accuracy and Precision than the other.

###### Advantages of Proposed System

* + - * More accurate result.
      * Able to detect cancer accurately behavior.
      * Cost and Time efficient.

# LITERATURE SURVEY

###### Breast Cancer Prediction by the Priyanka Gandhi and Prof. ShaliniL.

In this paper, ML techniques are explored in order to boost the accuracy of diagnosis.Methods such as CART, Random Forest, K-NearestNeighbours are compared. The dataset used is acquiredfrom UC Irvine Machine Learning Repository. It is foundthat KNN algorithm has much better performance than the other techniques used in comparison. The most accurate model was K-Nearest Neighbour. The classification modelsuch as Random Forest and Boosted Trees showed the similar accuracy. Therefore, the most accurate classier canbe used to detect the tumour so that the cure can be foundin early stage.

###### Breast Cancer Prediction Using Genetic Algorithm

Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach written by Pragya Chauhan and Amit Swami proposed a system where they found that Breastcancer prediction is an open area of research. In this paper dierent machine learning algorithms are used for detection of Breast Cancer Prediction. Decision tree, random forest,support vector machine, neural network, linear model,adabost, naive bayes methods are used for prediction.

###### S.Sidhu As investigated the performance of Support Vector Machine

S.Sidhu As investigated the performance of Support Vector Machine, Artificial Neural Network and Naïve Bayes using the Wisconsin Diagnostic Breast Cancer (WDBC) Dataset by integrating these machine leaning techniques with feature selection/feature extraction methods to obtain the most suitable one. The simulation results showed that SVM-LDA was chosen over all the other methods due to their longer computational time.

**SYSTEM REQUIREMENTS AND SPECIFICATIONS**

### GENERAL DESCRIPTION OF THE SYSTEM

A proper and thorough literature survey concludes that there are various methods that can be used to detect Breast cancer Prediction. Some of these approaches are:

1. Artificial Neural Network
2. Bayesian Network
3. Decision Trees
4. Random Forest Model
5. Supporting Vector Machine

#### Overview of Functional Requirements Preprocess Data

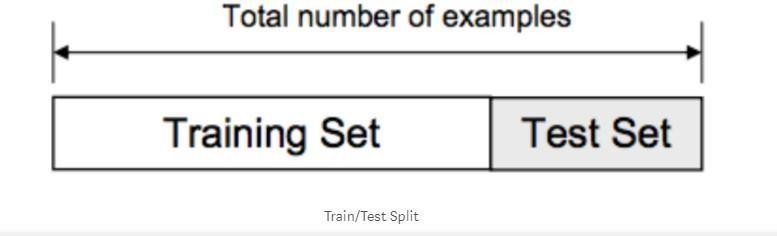
Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Therefore, certain steps are executed to convert the data into a small clean data set. This technique is performed before the execution of Iterative Analysis. The set of steps is known as Data Preprocessing. It includes,

* + - * Data Cleaning
      * Data Integration
      * Data Transformation
      * Data Reduction

#### Train and Test Data Creation

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later. We have the test dataset (or subset) in order to test our model’s prediction on this subset.



*Fig 3.1 Training and Test Dataset*

#### Model Creation

The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The term ML model refers to the model artifact that is created by the training process.

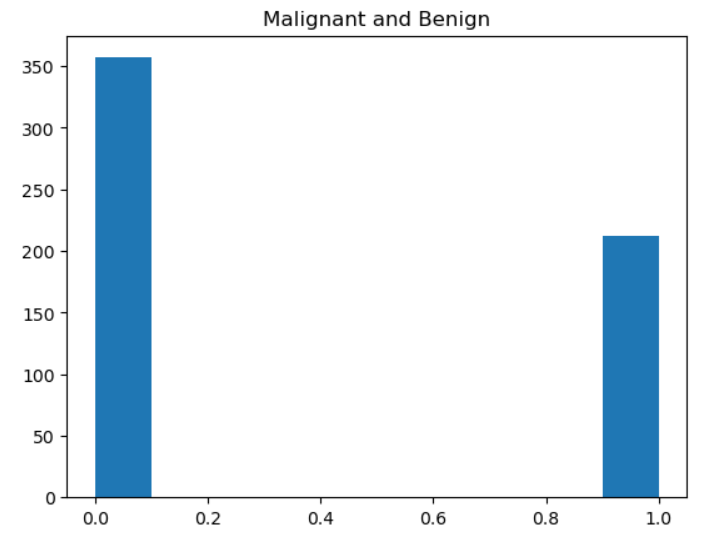
The training data must contain the correct answer, which is known as a target or target attribute. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict), and it outputs an ML model that captures these patterns.

You can use the ML model to get predictions on new data for which you do not know the target. For example, let's say that you want to train an ML model to predict if an email is spam or not spam. You would provide training data that contains emails for which you know the target (that is, a label that tells whether an email is spam or not spam). Machine would train an ML model by using this data, resulting in a model that attempts to predict whether new email will be spam or not spam.

In Our project we are using Supporting vector machine to build our model on breast Cancer dataset.

#### Result Analysis

In this final phase, we will test our model on our prepared dataset and measure the malignant or benign performance on our dataset. To evaluate the performance of our created classification and make it comparable to current approaches, we use Accuracy to measure the effectiveness of classifiers. We consider the Malignant as a negative class and Benign class as a positive class.



### TECHNICAL REQUIREMENTS OF THE SYSTEM

#### Hardware Requirements

System Processor : Core i3 / i5

Hard Disk : 500 GB.

Ram : 4 GB.

*Any desktop / Laptop system with above configuration or higher level.*

#### Software Requirements

Operating system : Windows 8 / 10 Programming Language : Python

Framework : Anaconda

IDE : Jupyter Notebook

DL Libraries : Numpy, Pandas, sklearn, matplot

### LANGUAGE SPECIFICATION

#### Python Introduction:

* + - * Python is an easy to learn, powerful programming language. It has efficient high-level datastructures and a simple but effective approach to object-oriented programming. Python’s elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms.
      * The Python interpreter and the extensive standard library are freely available in source or binary form for all major platforms from the Python Web site, *https://[www.python.org/](http://www.python.org/)*[,](http://www.python.org/) and may be freely distributed. The same site also contains distributions of and pointers to many free third party Python modules, programs and tools, and additional documentation.
      * The Python interpreter is easily extended with new functions and data types implemented in C or C++ (or other languages callable from C). Python is also suitable as an extension language for customizable applications.
      * Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.
      * Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
      * Python is Interactive − you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
      * Python is Object-Oriented − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
      * Python is a Beginner's Language − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

#### Machine Learning Features

Machine Learning is a method of statistical learning where each instance in a dataset is described by a set of features or attributes. In contrast, the term “Deep Learning” is a method of statistical learning that extracts features or attributes from raw data. Deep Learning does this by utilizing neural networks with many hidden layers, big data, and powerful computational resources. The terms seem somewhat interchangeable, however, with Deep Learning method, The algorithm constructs representations of the data automatically. In contrast, data representations are hard-coded as a set of features in machine learningalgorithms, requiring further processes such as feature selection and extraction, (such as PCA).

Both terms are in dramatic contrast with another class of classical artificial intelligence algorithmsKnown as Rule-Based Systems where each decision is manually programmed in such a way that it resembles a statistical model.

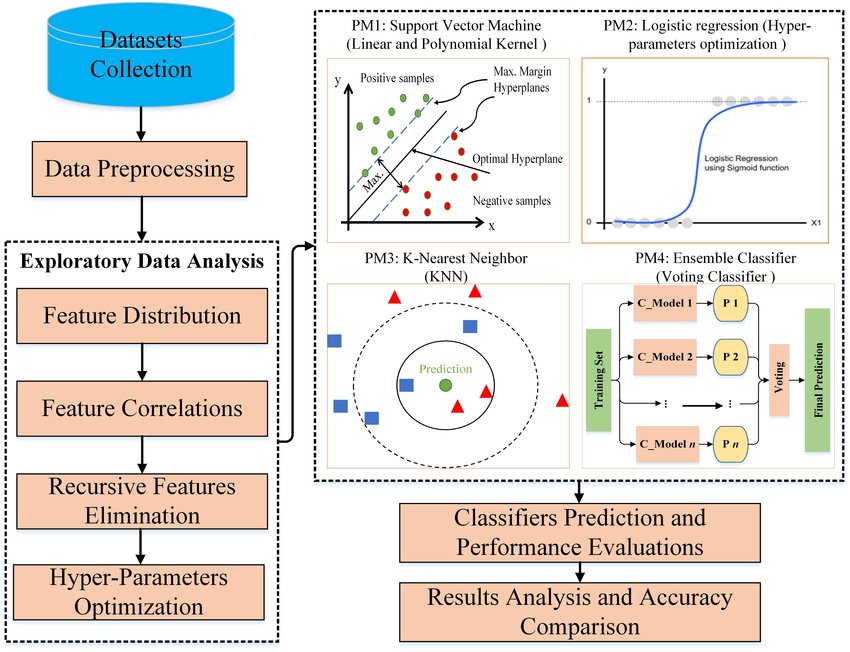
In Machine Learning and Deep Learning, there are many different models that fall into two different categories, supervised and unsupervised. In unsupervised learning, algorithms such as k-Means, hierarchical clustering, and Gaussian mixture models attempt to learn meaningful structures in the data. Supervised learning involves an output label associated with each instance in the dataset. This outputcan be discrete/categorical or real-valued. Regression models estimate real-valued outputs, whereas classification models estimate discrete-valued outputs. Simple binary classification models have just two output labels, 1 (positive) and 0 (negative). Some popular supervised learning algorithms that are considered Machine Learning: are linear regression, logistic regression, decision trees, random forest, support vector machines, and neural networks, as well as non-parametric models such as k-Nearest Neighbors.

**Support Vector Machine (SVM)**

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. The objective is to find aplane that has the maximum margin,i.e the maximum distance between data points of both classes.

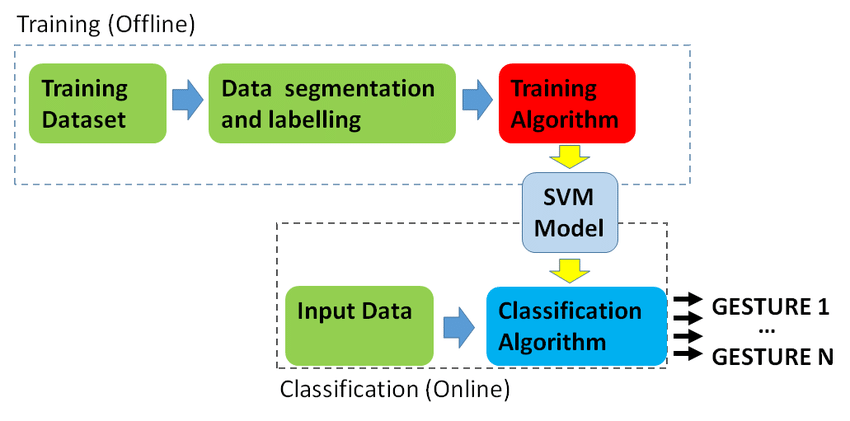
**SYSTEM DESIGN AND ANALYSIS**

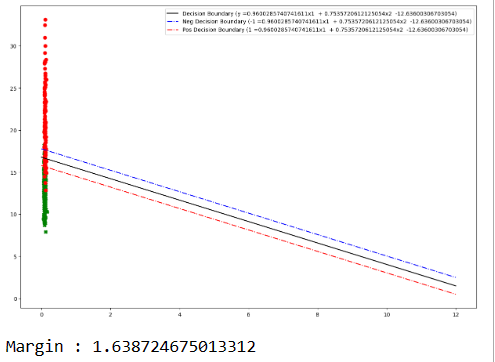
### SYSTEM ARCHITECTURE

****

*Fig 4.1 System Architecture*

### BLOCK DIAGRAM

**SVM HYPERPLANE:**

****

# CHAPTER-5

# IMPLEMENTATION

In this work, a business intelligent model has been developed, to classify different animals, based on a specific business structure deal with Animal classification using a suitable machine learning technique. The model was evaluated by a scientific approach to measure accuracy. We are using Supportng vector machine to build our model.

**CODE:**

#Import necessary packages import pandas as pd

from sklearn import svm

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score,confusion\_matrix

import seaborn as sns

import matplotlib.pyplot as plt

# Load the csv file

data=pd.read\_csv("brca.csv")

data

#Checking dataframe for null values

data.isnull().any()

#Plot the data count

x=data.drop(['y'],axis=1)

y=data['y']

plt.subplot(2,1,2)

plt.plot(x,y)

#Splitting the dataset

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,random\_state=42,test\_size=0.2)

print(len(x\_train),len(y\_train)

#Create a SVM model

s\_v\_m=svm.SVC(kernel="linear")

s\_v\_m.fit(x\_train,y\_train)

#Create an function called “metrics” to display the metrics

from sklearn.metrics import precision\_score,recall\_score,f1\_score

acc\_test=s\_v\_m.predict(x\_test)

acc1=accuracy\_score(acc\_test,y\_test)

print(acc1)

acc\_train=s\_v\_m.predict(x\_train)

acc1=accuracy\_score(acc\_train,y\_train)

print(acc1)

print("Precision Score: ", precision\_score(y\_test,acc\_test, average="macro").mean()\*100)

print("Recall Score: ", recall\_score(y\_test,acc\_test , average="macro").mean()\*100)

print("F1 Score: ",f1\_score(y\_test,acc\_test, average="macro").mean()\*100

#Checking the trained model

import numpy as np

input\_data=(24.01,1.0,85.63,5,0.1075,0.127,0.04568,0.0311,0.7,0.06811,0.1852,0.7477,1.383,14.67,0.004097,0.01898,0.01698,0.00649,0.01678,0.25,14.5,20.49,96.09,630.5,0.12,0.2776,0.189,0.07283,0.3184,0.83)

inputas=np.asarray(input\_data)

input\_datare=inputas.reshape(1,-1)

prediction3=s\_v\_m.predict(input\_datare)

if (prediction3==0):

print("it is benign")

else:

print("it is malignant")

#Checking the trained model

import numpy as np

input\_data=(24.01,1.0,85.63,5,0.1075,0.127,0.04568,0.0311,0.7,0.06811,0.1852,0.7477,1.383,14.67,0.004097,0.01898,0.01698,0.00649,0.01678,0.25,14.5,20.49,96.09,630.5,0.12,0.2776,0.189,0.07283,0.3184,0.83)

inputas=np.asarray(input\_data)

input\_datare=inputas.reshape(1,-1)

prediction3=s\_v\_m.predict(input\_datare)

if (prediction3==0):

print("it is benign")

else:

print("it is malignant")

#Create another model for increse accuracy:

from sklearn.model\_selection import GridSearchCV

param\_grid = {'C': [0.1,100,1000,10000],

'gamma':[0.01,0.001,0.0001,0.00001],

'kernel': ['rbf']}

bc\_grid = GridSearchCV(s\_v\_m, param\_grid, refit = True, verbose = 3)

bc\_grid.fit(x\_train, y\_train)

#Checking the Gridsearch cv

predict=bc\_grid.predict(x\_test)

acc=accuracy\_score(y\_test,predict)

print(acc)

#creating hyperplane

import matplotlib.pyplot as plt

from sklearn.svm import SVC

X=data.get(['smoothness\_mean','radius\_worst']).to\_numpy()

Y=data['y']

clf=svm.SVC( gamma='auto',kernel='linear')

clf.fit(X,Y)

# plot the decision boundary ,data points,support vector etcv

w = clf.coef\_[0]

a = -w[0] / w[1]

xx = np.linspace(0,12)

yy = a \* xx - clf.intercept\_[0] / w[1]

y\_neg = a \* xx - clf.intercept\_[0] / w[1] + 1

y\_pos = a \* xx - clf.intercept\_[0] / w[1] - 1

plt.figure(1,figsize= (15, 10))

plt.plot(xx, yy, 'k',

label=f"Decision Boundary (y ={w[0]}x1 + {w[1]}x2 {clf.intercept\_[0] })")

plt.plot(xx, y\_neg, 'b-.',

label=f"Neg Decision Boundary (-1 ={w[0]}x1 + {w[1]}x2 {clf.intercept\_[0] })")

plt.plot(xx, y\_pos, 'r-.',

label=f"Pos Decision Boundary (1 ={w[0]}x1 + {w[1]}x2 {clf.intercept\_[0] })")

for i in range(500):

if (Y[i]==1):

plt.scatter(X[i][0], X[i][1],color='red', marker='o')

elif (Y[i]==0):

plt.scatter(X[i][0], X[i][1],color='green', marker='X')

plt.legend()

plt.show()

# calculate margin

print(f'Margin : {2.0 /np.sqrt(np.sum(clf.coef\_ \*\* 2)) }')

### DIFFERENT MODULES OF THE PROJECT

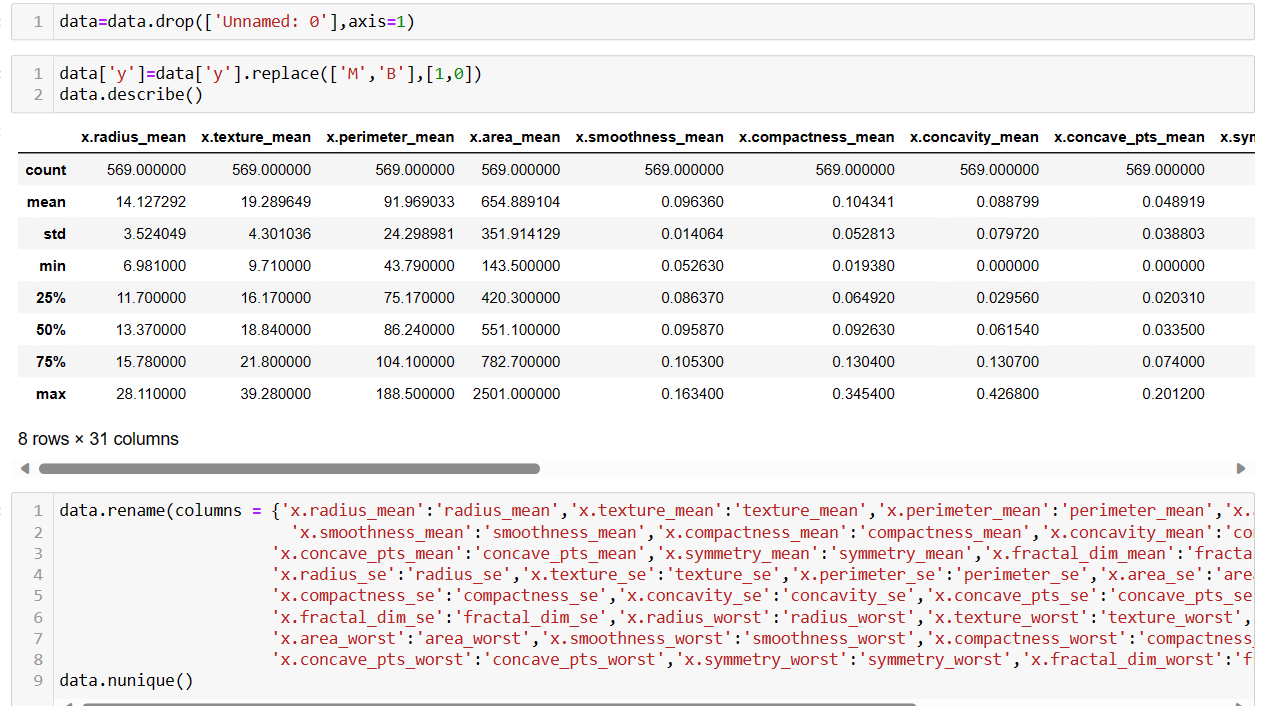
1. Perform Exploratory Data Analysis (EDA)
2. Apply different Machine Learning algorithms

1. Supporting vector machine

1. Train and evaluate our models on the dataset and pick the best

## Perform Exploratory Data Analysis (EDA) on the Dataset

There are a total of 284,807 transactions with only 492 of them being fraud. check for null values and analyze the dataset



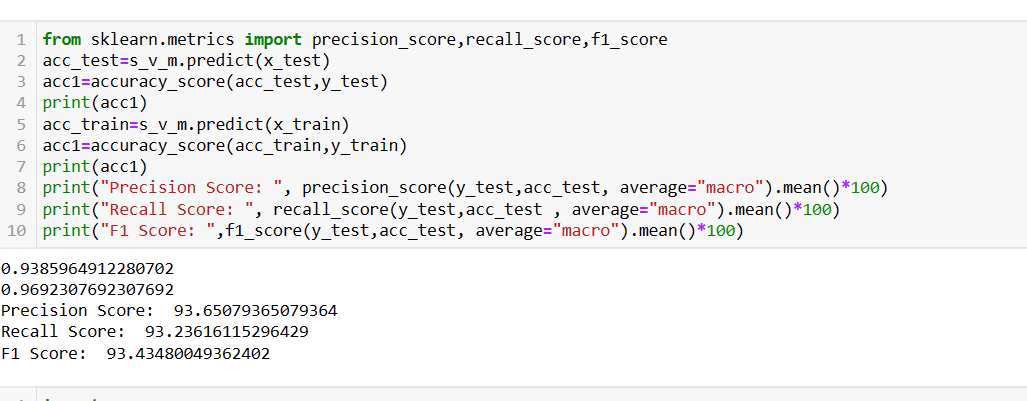
*Fig 5.3 Perform EDA on the Dataset*

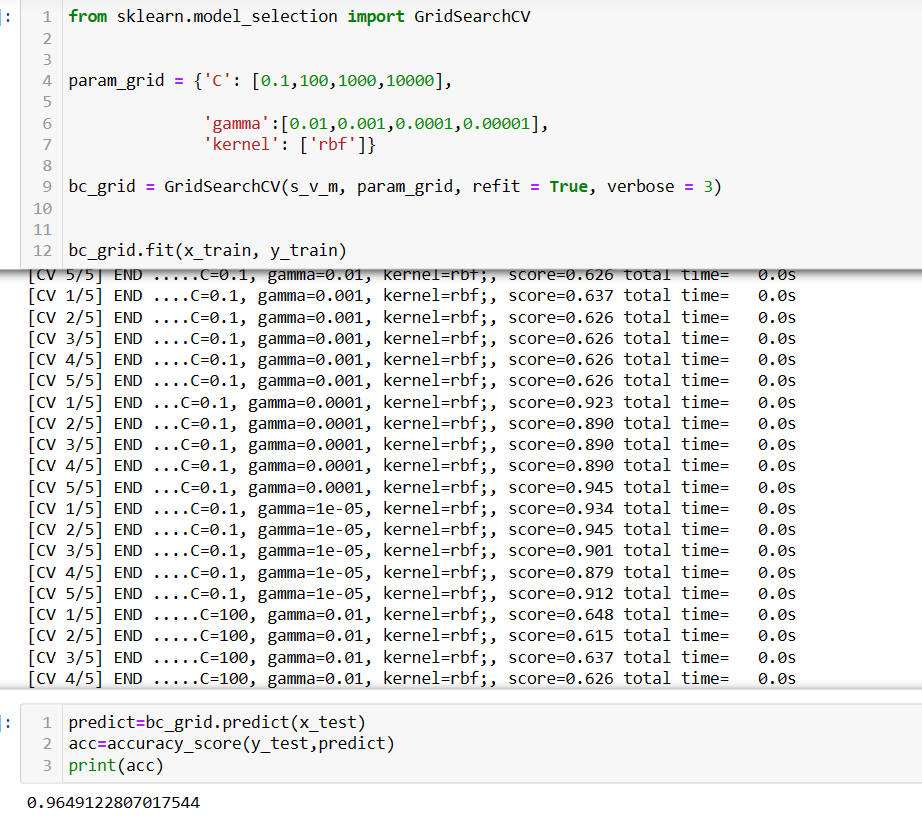
## Apply different Machine Learning algorithms to our dataset

**Supporting Vector machine**



## Train and evaluate our models on the dataset and pick the best





**CONCLUSION AND FUTURE ENHANCEMENT**

### CONCLUSION

Breast cancer if found at an early stage will help save lives of thousands of women or even men. These projects help the real world patients and doctors to gather as much information as they can. The research on nine papers has helped us gather the data for the project proposed by us. By using machine learning algorithms we will be able to classify and predict the cancer into being or malignant.

### FUTURE ENHANCEMENT

The project has covered almost all the requirements. Further requirements and improvements can easily be done since the coding is mainly structured or modular in nature. Improvements can be appended by changing the existing modules or adding new modules. One important development that can be added to the project in future is analyzing and creating own algorithm as like SVM to improve our accuracy.

By doing that , the chance of getting more accurate result will be higher than this.

**REFERENCES**

1. [Support Vector Machine (SVM) Algorithm - GeeksforGeeks](https://www.geeksforgeeks.org/support-vector-machine-algorithm/)
2. [Prediction of Breast Cancer using Machine Learning Approaches - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9175124/)
3. [How to Identify Overfitting Machine Learning Models in Scikit-Learn - MachineLearningMastery.com](https://machinelearningmastery.com/overfitting-machine-learning-models/)
4. [GitHub - emirhanai/Classification-thanks-to-the-SVM-model-with-7-years-of-ozone-data-with-Machine-Learning!](https://github.com/emirhanai/Classification-thanks-to-the-SVM-model-with-7-years-of-ozone-data-with-Machine-Learning)
5. [Breast Cancer Wisconsin (Diagnostic) Data Set (kaggle.com)](https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data)
6. <https://www.wps.com/academy/how-to-delete-a-word-doc-page-quick-tutorials-1864265/>
7. <https://www.researchgate.net/figure/Breast-cancer-detection-system-architecture_fig1_356115197>
8. <https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/auto_examples/svm/plot_separating_hyperplane.html>
9. <https://www.geeksforgeeks.org/svm-hyperparameter-tuning-using-gridsearchcv-ml/>
10. <https://stackoverflow.com/questions/34113083/numpy-contour-typeerror-input-z-must-be-a-2d-array>