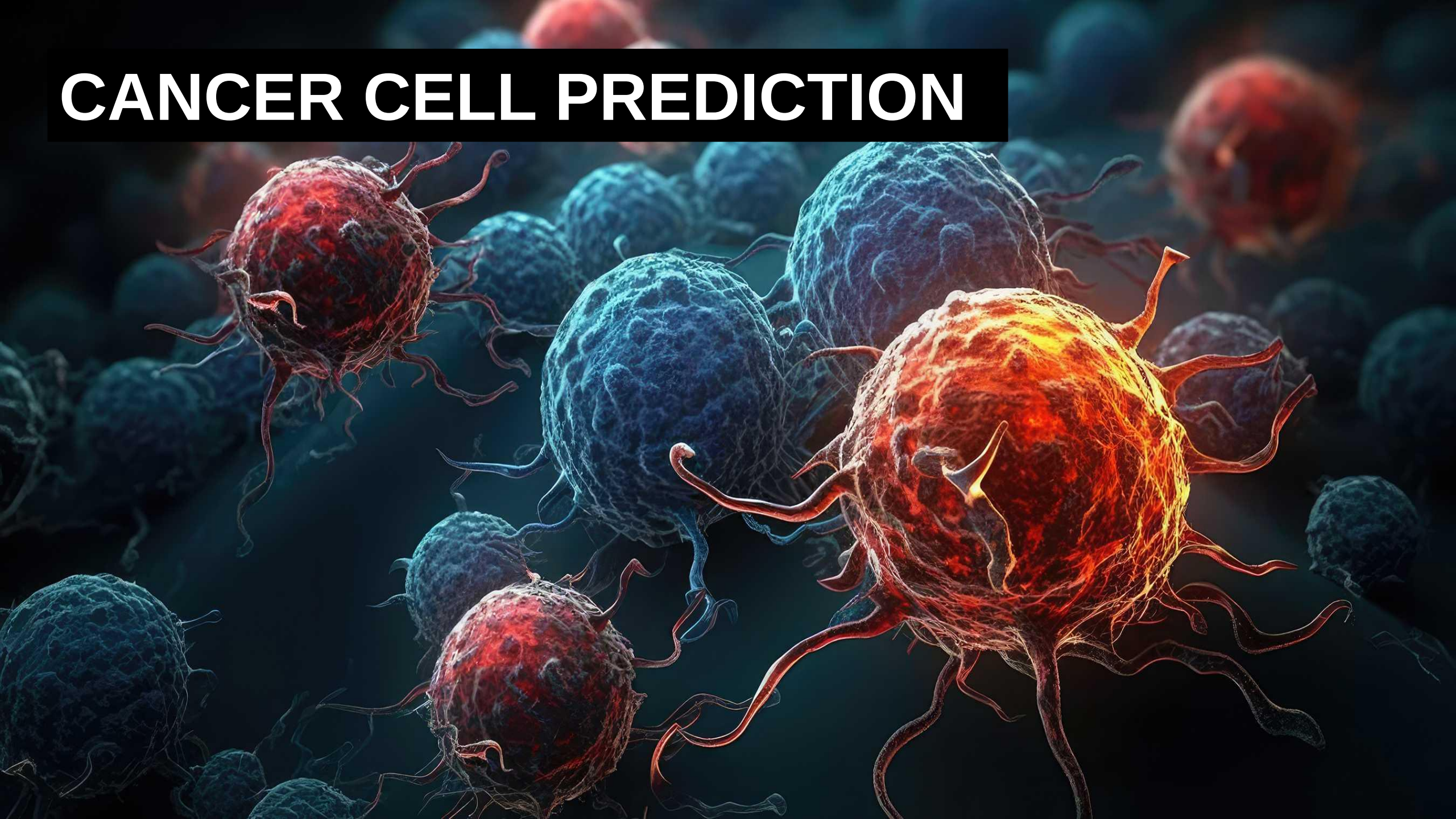
CANCER CELL PREDICTION



CONTENTS

INTRODUCTION

Breast cancer stands as a predominant cause of mortality among women globally, necessitating effective tools for its prediction and early diagnosis. The complexity of medical data analysis makes the prediction of breast cancer a challenging task. The integration of **machine learning (ML) algorithms** has emerged as a valuable solution in assisting doctors and pathologists in decision-making processes and distinguishing between malignant and benign tumors. This study explores the application of Support Vector Machines (SVMs), a powerful ML algorithm, in breast cancer prediction.

Research indicates that ML techniques play a pivotal role in decision-making processes related to breast cancer prediction. By leveraging data mining techniques, the study proposes a method to reduce the reliance on conventional tests, such as MRI, mammogram, ultrasound, and biopsy, by focusing on detecting the presence of the risk of breast cancer. The proposed method utilizes a dataset available in the sklearn library, consisting of unique ID numbers, corresponding diagnoses (malignant/benign), and real-value features (parameters).

ABSTRACT

SOFTWARE DEVELOPMENT KIT :

An SDK (software development kit) is a collection of tools for developing applications for specific hardware/software or in a certain programming language. With some interpreted languages, the SDK can be identical to the run-time environment. SDKs typically include an integrated development environment (IDE), which serves as the central programming interface.

The "google.colab" library provides functionality for tasks such as 4 importing and exporting files, installing Python packages, managing Colab sessions, and connecting to external services like Google Drive and Google Sheets. Some of the common tasks that can be performed using the "google.colab"

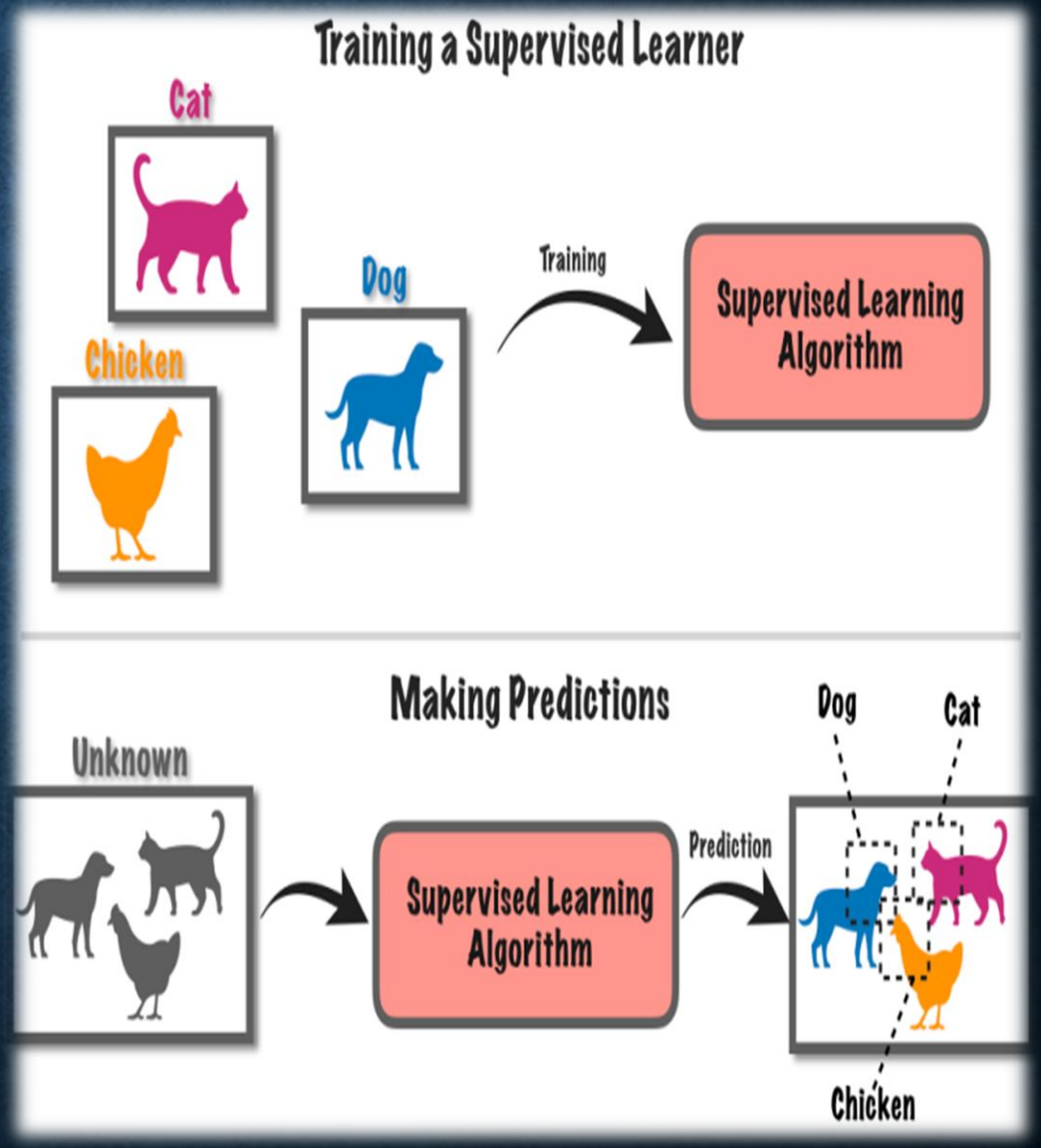
LIBRARIES USED :

- ✓ Numpy
- ✓ Pandas
- ✓ Matplotlib
- ✓ Pyplot
- ✓ Seaborn

SUPERVISED LEARNING

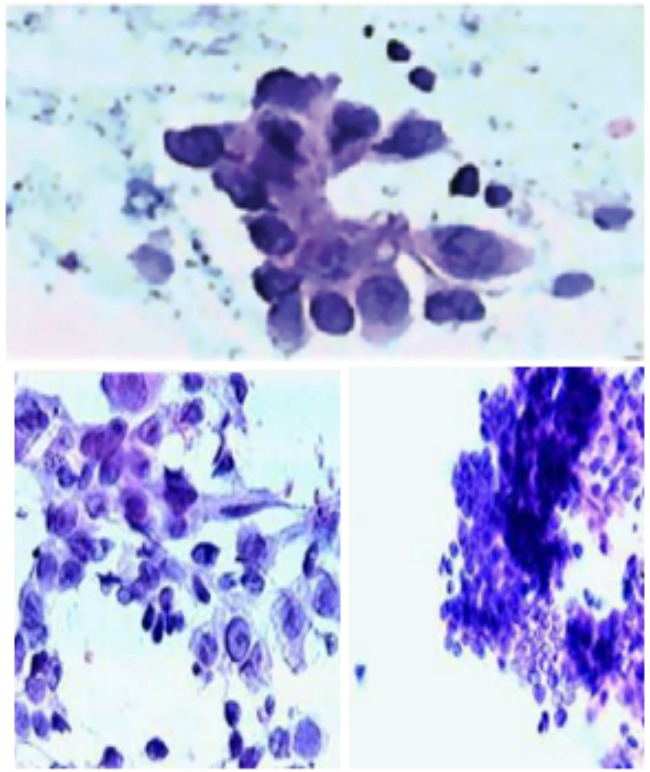
ML encompasses a broad range of tasks and methods. Supervised learning tasks have a known available outcome to predict, such as presence of a tumour, length of survival, or treatment response. Unsupervised learning identifies patterns and subgroups within data where there is no clear outcome to predict. It is often used for more exploratory analysis.

Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically, supervised learning is when we teach or train the machine using data that is well-labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labelled data.



Breast Cancer Detection using Machine Learning

Breast Tumor Images



Extracted Features

- Radius
- Texture
- Area
- Perimeter
- Smoothness
- Compactness
- Concavity
- Symmetry
- Fractal Dimension

Data pre-processing

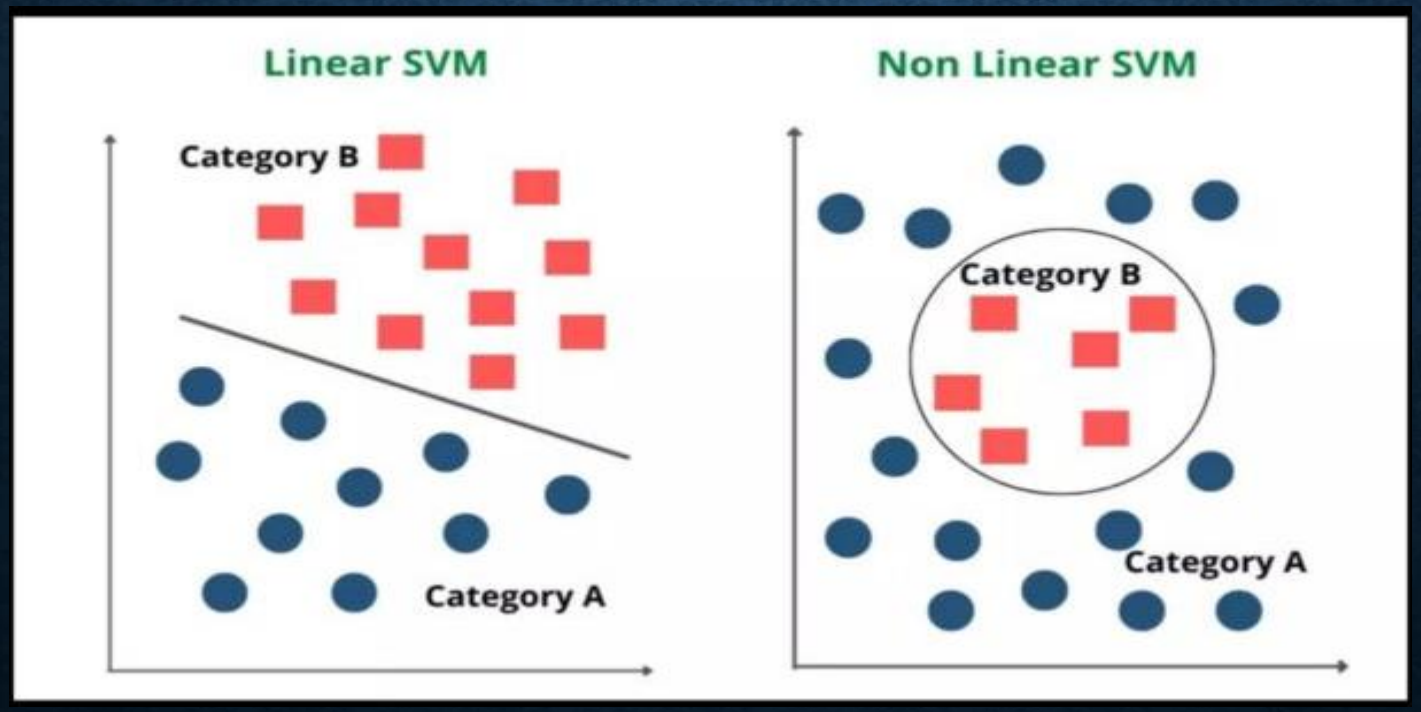
Machine Learning

Classification

- Malignant
- Benign

ALGORITHM USED

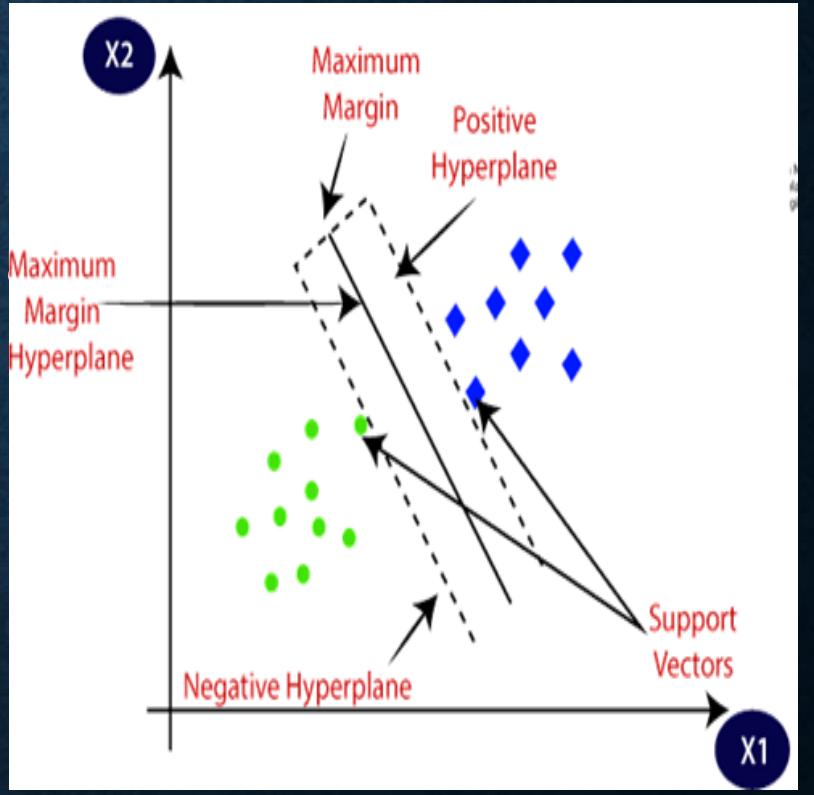
SUPPORT VECTOR MACHINE



SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) is a powerful machine learning algorithm used for linear or nonlinear classification, regression, and even outlier detection tasks.

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. SVMs can be used for a variety of tasks, such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection. The main objective of the SVM algorithm is to find the optimal hyperplane in an N- dimensional space that can separate the data points in different classes in the feature space. The hyperplane tries that the margin between the closest points of different classes should be as maximum as possible. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three. SVMs are adaptable and efficient in a variety of applications because they can manage high-dimensional data and nonlinear relationships.



❑ IMPLEMENTING SVM IN PYTHON

SVM KERNELS

- SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions. These functions can be different types. For example **linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid**.

- The most used type of kernel functions **RBF**. Because it has localized and finite response along the entire x-axis. The kernel functions return the inner product between two points in a suitable feature space. Thus by defining a notion of similarity, with little computational cost even in very high-dimensional spaces.

- **Kernel Function** is a method used to take data as input and transform it into the required form of processing data. “Kernel” is used due to a set of mathematical functions used in Support Vector Machine providing the window to manipulate the data. So, Kernel Function generally transforms the training set of data so that a nonlinear decision surface is able to transform to a linear equation in a higher number of dimension spaces.

- Polynomial $K(a,b) = (1 + \sum_j a_j b_j)^d$
- Radial Basis Functions

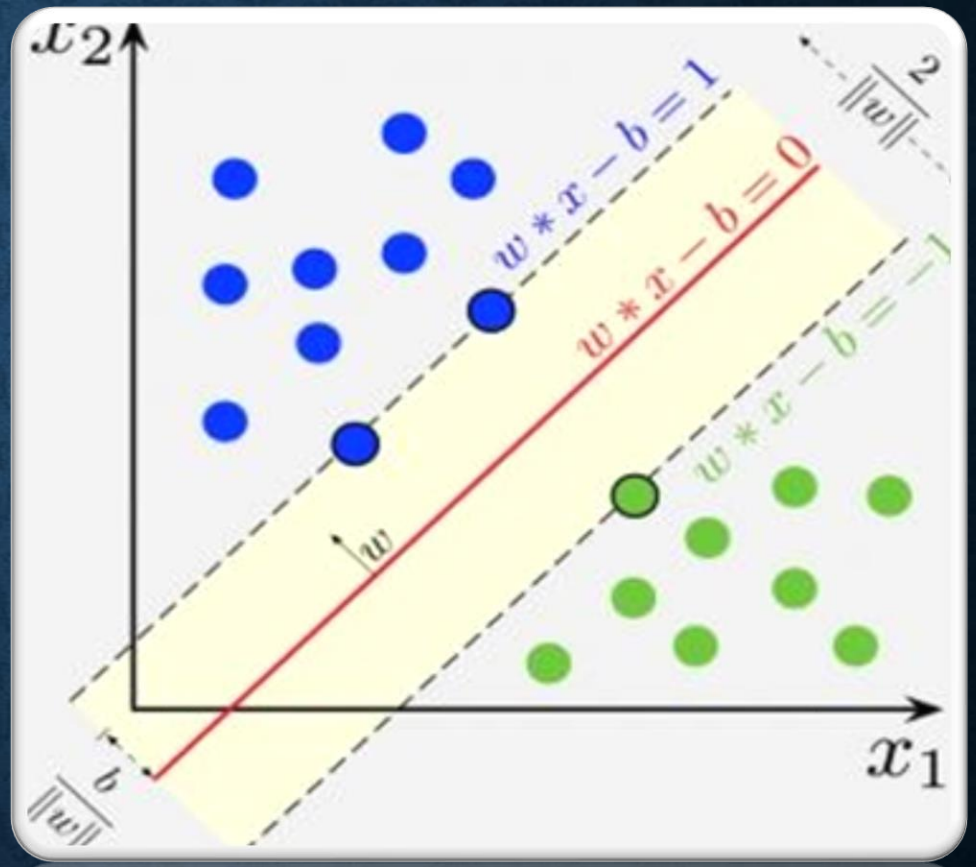
$$K(a,b) = \exp(-(a-b)^2/2\sigma^2)$$

- Saturating, sigmoid-like:

$$K(a,b) = \tanh(ca^T b + h)$$

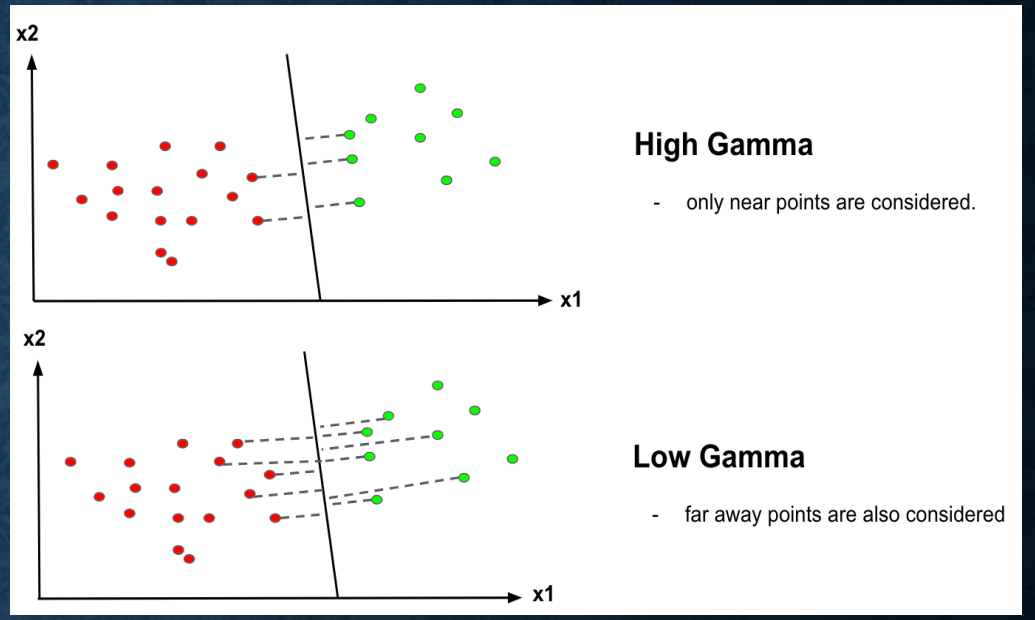
SVC

An SVM classifier, or support vector machine classifier, is a type of machine learning algorithm that can be used to analyze and classify data. A support vector machine is a supervised machine learning algorithm that can be used for both classification and regression tasks. The Support vector machine classifier works by finding the hyperplane that maximizes the margin between the two classes. The Support vector machine algorithm is also known as a max-margin classifier. Support vector machine is a powerful tool for machine learning and has been widely used in many tasks such as hand-written digit recognition, facial expression recognition, and text classification.



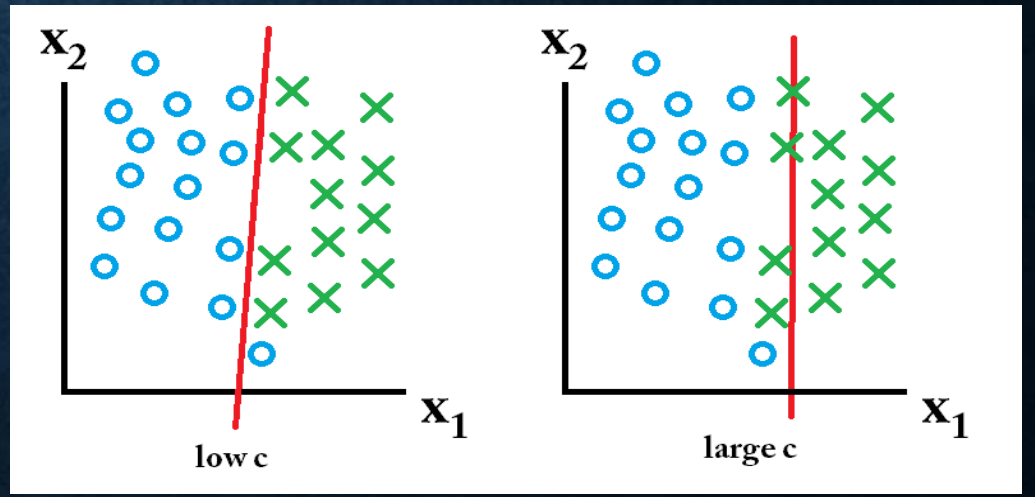
GAMMA

- Gamma is used when we use the Gaussian RBF kernel.
- Gamma is a hyper parameter which we have to set before training model.
- Gamma decides that how much curvature we want in a decision boundary.
- Gamma high means more curvature. • Gamma low means less curvature.

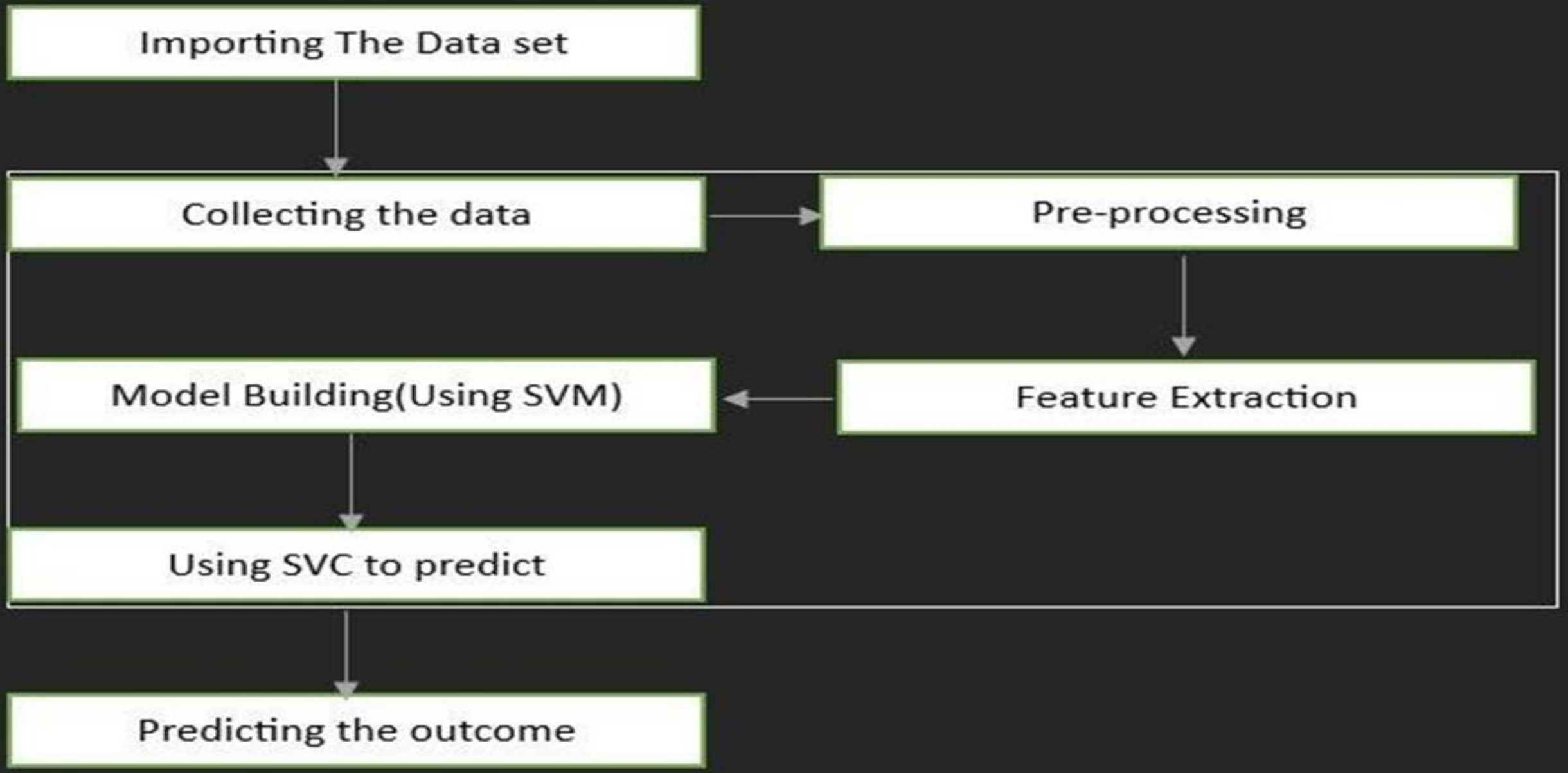


C

- **C** is a hypermeter which is set before the training model and used to control error



WORK FLOW OF THE PROJECT



❑ SOURCE CODE & OUTPUT

[1]
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.datasets import load_breast_cancer
```

```
cancer=load_breast_cancer()
x=cancer.data[:,:2]
y=cancer.target
y_name=cancer.target_names

xd=pd.DataFrame(x)
yd=pd.DataFrame(y_name)

print(y)
print(xd)
print(x)
```

```
[0 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 0 0 0
 1 0 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 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 0 1 0 1 1 0 1 1 0 1 0 0 1 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 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 0 1 1 0 1 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 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 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 0 0 0 0 0 0 0 1]
          0      1
0    17.99  10.38
1    20.57  17.77
2    19.69  21.25
3    11.42  20.38
4    20.29  14.34
..     ...    ...
564  21.56  22.39
```

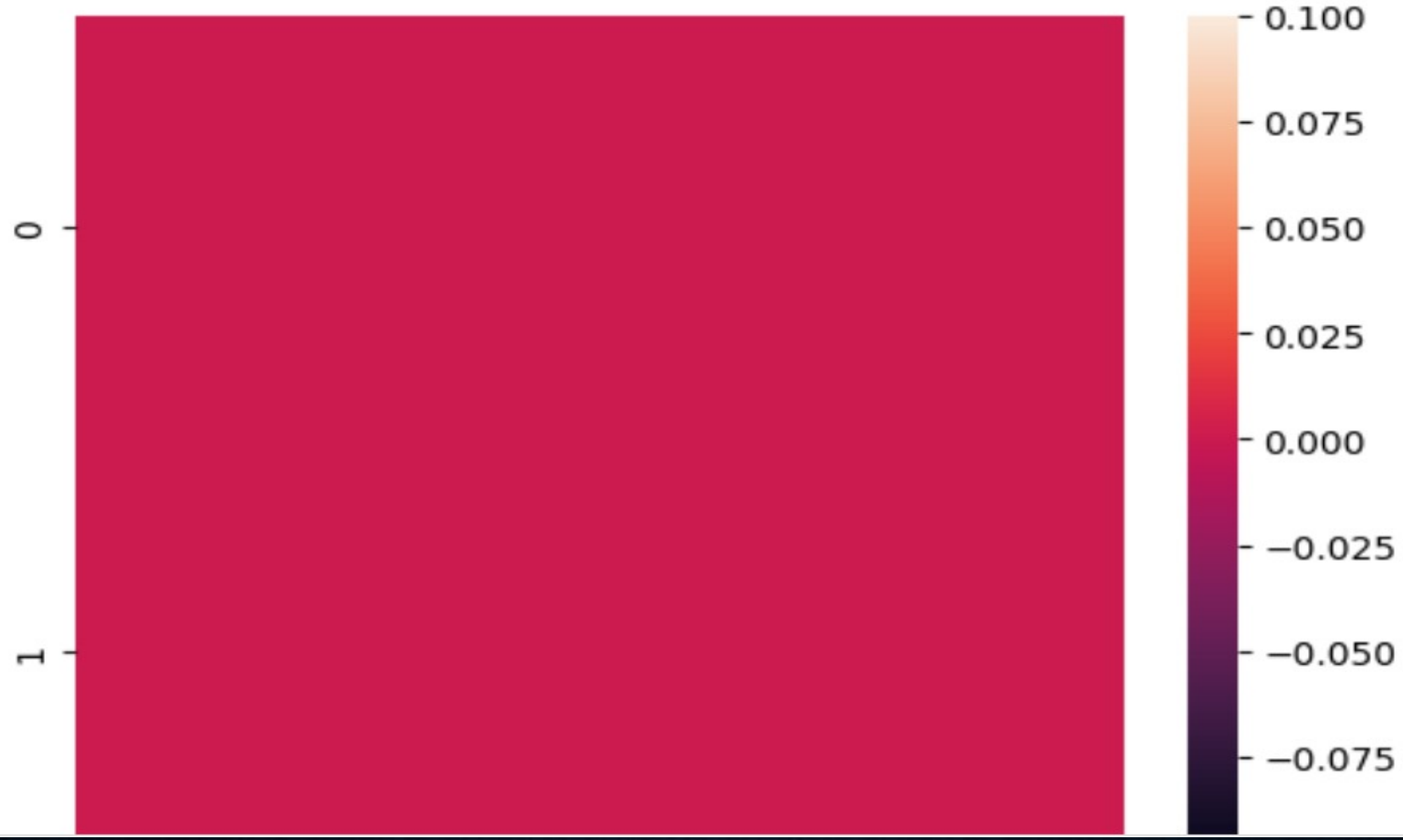
```
..      ...     ...
564   21.56   22.39
565   20.13   28.25
566   16.60   28.08
567   20.60   29.33
568    7.76   24.54

[569 rows x 2 columns]
[[17.99 10.38]
 [20.57 17.77]
 [19.69 21.25]
 ...
 [16.6  28.08]
 [20.6  29.33]
 [ 7.76 24.54]]
```

[3]
```
sns.heatmap(yd.isnull())
from sklearn.inspection import DecisionBoundaryDisplay
from sklearn.svm import SVC
svm_model=SVC(kernel="rbf",gamma=.5,C=1)
svm_model.fit(x,y)
```

SVC

```
SVC(C=1, gamma=0.5)
```

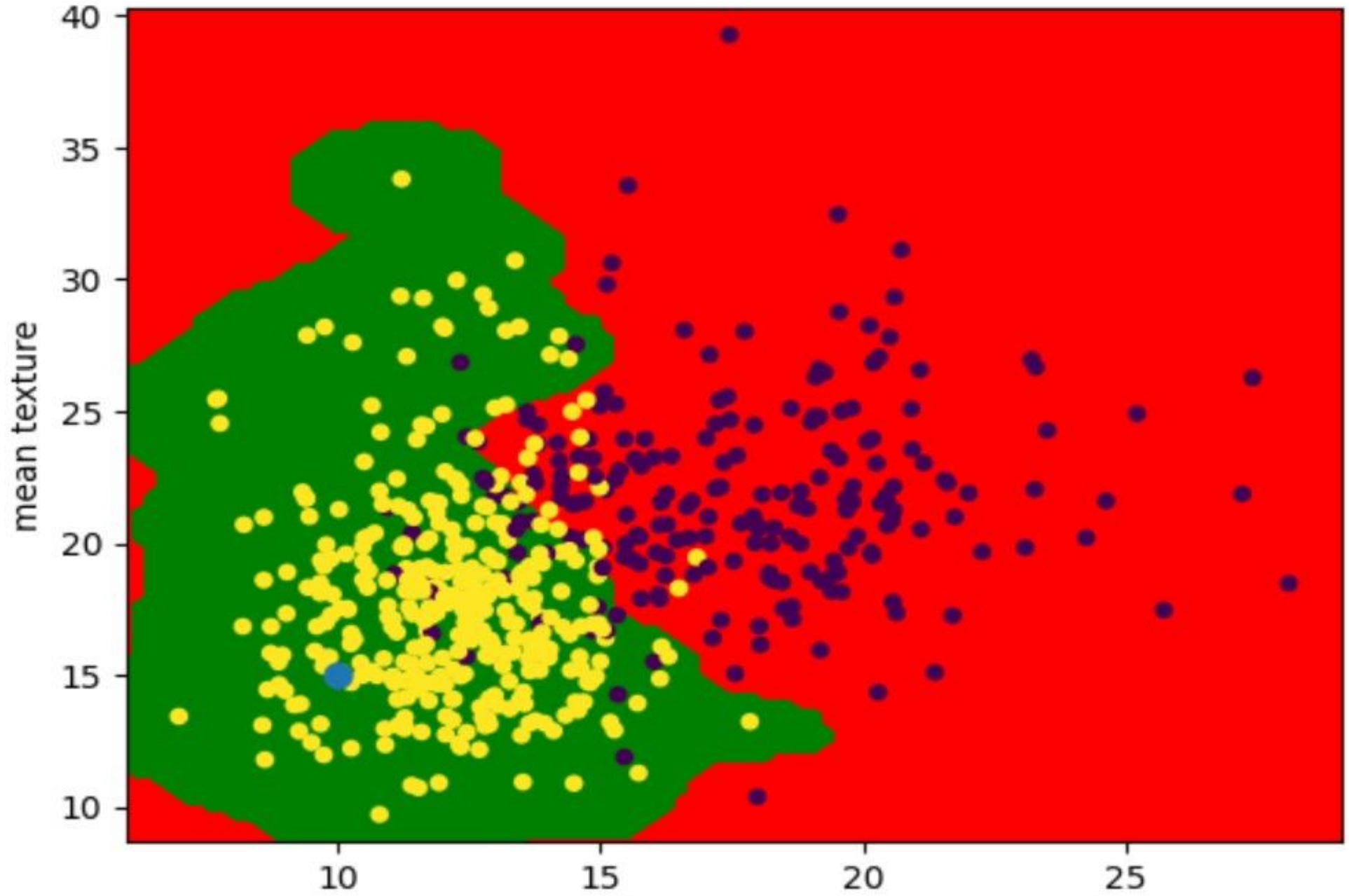


```
import matplotlib.colors
mycol=matplotlib.colors.ListedColormap(["red","green"])
DecisionBoundaryDisplay.from_estimator(
    svm_model,
    x,
    response_method="predict",
    xlabel=cancer.feature_names[0],
    ylabel=cancer.feature_names[1],
    cmap=mycol

)
radius=x[:,0]
txt=x[:,1]
plt.scatter(radius,txt,c=y,s=20)
plt.scatter(10,15,s=60)
inp=[[10,15]]
yp=svm_model.predict(inp)
print(y_name[yp[0]])
```

```
benign
```

benign



✓ 1s completed at 3:25 PM

LIMITATION

Predicting cancer cell behavior using machine learning techniques has shown promise in various areas such as diagnosis, prognosis, and treatment response prediction. However, there are several limitations and challenges associated with cancer cell prediction in machine learning:

- Machine learning models require large amounts of high-quality data for training. Obtaining large, diverse, and well-annotated datasets in cancer research can be challenging due to issues such as data availability, data heterogeneity, and data labeling errors.

- Cancer datasets often contain a large number of features (genes, proteins, etc.), many of which may be irrelevant or redundant. Selecting the most informative features and reducing dimensionality while retaining relevant information is crucial for building accurate and interpretable models.

- Overfitting occurs when a model learns to capture noise or random fluctuations in the training data rather than the underlying patterns. This can lead to poor generalization performance when the model is applied to unseen data. Regularization techniques and cross-validation strategies are commonly used to mitigate overfitting.

- Some machine learning models, such as deep neural networks, are often considered "black-box" models because they lack interpretability, making it challenging to understand the underlying reasons behind model predictions. Interpretable models or post-hoc interpretation methods are needed to provide insights into the factors driving cancer cell predictions.

- Imbalanced datasets can bias model training and evaluation, leading to suboptimal performance, particularly for minority classes.

CONCLUSION

Breast cancer is the important field of research and technology helps to reduce mortality rate caused by breast cancer. There are many ML algorithms introduced till now for analysis of medical datasets. It is essential for a medical diagnosis that the data on breast cancer be classified in a way that is both accurate and effective. Even though many numbers of methods have been developed to classify breast cancer data, there are still many obstacles to overcome, including accuracy. In order to address thisissue, we put forth a model for the classification of data relating to breast cancer. In this paper we applied SVM, ML Classification technique. The proposed machine-learning approaches could predict breast cancer as the early detection of this disease could help slow down the progress of the disease and reduce the mortality rate through appropriate therapeutic interventions at the right time. Applying different machine learning approaches, accessibility to bigger datasets from different institutions (multi centre study) and considering key features from a variety of relevant data sources could improve the performance of modelling. In conclusion, the implementation of SVM can produce almost enough accuracy to be termed as a medically acceptable level of diagnostic accuracy for the dataset used. However, the dataset is not highly normalized resulted in an over-fitting problem due to the number of prominent outliers as seen while tuning the parameters. This limitation partly affected the accuracy of the model.

FUTURE SCOPE

AI is set to change the medical industry in the coming decades — it wouldn't make sense for pathology to not be disrupted too. Currently, ML models are still in the testing and experimentation phase for cancer prognoses. As datasets are getting larger and of higher quality, researchers are building increasingly accurate models. Here's what a future cancer biopsy might look like: You perform clinical tests, either at a clinic or at home. Data is inputted into a pathological ML system. A few minutes later, you receive an email with a detailed report that has an accurate prediction about the development of your cancer. While you might not see AI doing the job of a pathologist today, you can expect ML to replace your local pathologist in the coming decades, and it's pretty exciting! ML models still have a long way to go, most models still lack sufficient data and suffer from bias. Yet, something we are certain of is that ML is the next step of pathology, and it will disrupt the industry.

BIBLIOGRAPHY

1. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8-17.

2 .Data Visualization Storytelling Using Data,written by Sharada Sringeswara,Purvi Tiwari,U. Dinesh Kumar.

3. Python Data Analytics(with pandas,numpy,matplotlib),written by Fabio Nelli.

4 . Data Analytics with Python,written by Dr. Bhavesh Devra,Dr. Dilip Kumar,Dr. Shajahan Basheer,Dr. Proloy Ghosh.

THANK YOU