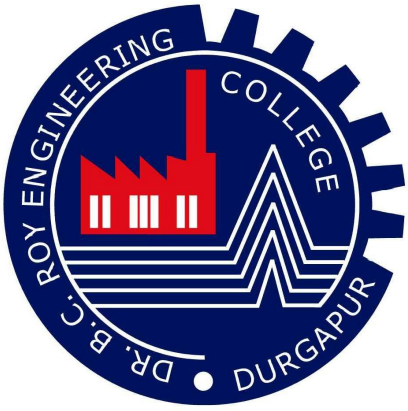
BREAST CANCER DETECTION USING MACHINE LEARNING



Raneet Roy

Abhinav Kumar

Avishek Kumar Bose

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DR. B.C. ROY ENGINEERING COLLEGE, DURGAPUR, WB

November, 2023

BREAST CANCER DETECTION USING MACHINE LEARNING

Report submitted to

Department of Computer Science and Engineering

Dr. B.C. Roy Engineering College, Durgapur, WB

for the partial fulfillment of the requirement to award the degree

of

Bachelor of Technology

in

Computer Science and Engineering

by

Raneet Roy 12000120018

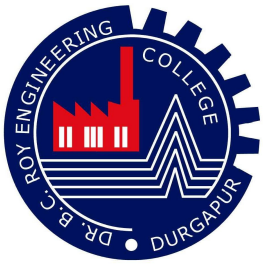
Abhinav Kumar 12000120015

Avishek Kumar Bose 12000120014

under the guidance

of

Supervisor: Prof. Suvobrata Sarkar



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DR. B.C. ROY ENGINEERING COLLEGE, DURGAPUR, WB

November, 2023

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DR. B.C. ROY ENGINEERING COLLEGE, DURGAPUR, WB

DECLARATION

We the undersigned, hereby declare that our B.Tech final year Project entitled, **"Breast Cancer Detection using Machine Learning"** is original and is our own contribution. To the best of out knowledge, the work has not been submitted to any other Institute for the award of any degree or diploma. We declare that we have not indulged in any form of plagiarism to carry out this project and/or writing this project report. Whenever we have used materials (data, theoretical analysis, figures, and text) from other sources, we have given due credit to them by citing in the text of the report and giving their details in the references. Finally, we undertake the total responsibility of this work at any stage here after.

Signature of the Students

_____

Raneet Roy (12000120018)

_____

Abhinav Kumar (12000120015)

_____

Avishek Kumar Bose (12000120014)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DR. B.C. ROY ENGINEERING COLLEGE, DURGAPUR, WB

RECOMMENDATION

This is to recommend that the work undertaken in this report entitled, **"Breast Cancer Detection using Machine Learning"** has been carried out by **"Raneet Roy, Abhinav Kumar, Avishek Kumar Bose"** under my/our supervision and guidance during the academic year 2023-24. This may be accepted in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (Computer Science and Engineering).

Prof. Suvobrata Sarkar	**Dr. Arindam Ghosh**
Assistant Professor,	Head of Department,
Department of CSE	Department of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DR. B.C. ROY ENGINEERING COLLEGE, DURGAPUR, WB

APPROVAL

This is to certify that, **Raneet Roy, Abhinav Kumar and Avishek Kumar Bose**, students in the Department of Computer Science & Engineering, worked on the project entitled **"Breast Cancer Detection using Machine Learning ".**

I hereby recommend that the report prepared by them may be accepted in partial fulfillment of the requirement of the Degree of Bachelors of Technology in the Department of Computer Science and Engineering, Dr. B.C. Roy Engineering College, Durgapur.

Examiners

_____

_____

_____

Prof. Suvobrata Sarkar

(Supervisor)

_____

Dr. Arindam Ghosh

(HOD, CSE)

_____

Date:

Prof. Biswadev Goswami

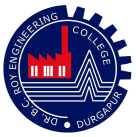
(Project Co-ordinator)

Place: DURGAPUR

_____

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DR. B.C. ROY ENGINEERING COLLEGE, DURGAPUR, WB



ACKNOWLEDGEMENT

It is our privilege to express our sincere regards to our project supervisor, Prof. Suvobrata Sarkar, for valuable inputs, able guidance, encouragement, whole-heartedcooperation, and constructive criticism throughout our project.

We deeply express our sincere thanks to the Head of Department, Dr.Arindam Ghosh, for encouraging and allowing us to present the project on the topic **"Breast Cancer Detection using Machine Learning"** at our department premises for partial fulfillment of the requirements leading to the award of the B.Tech. Degree.

Furthermore, we would also like to acknowledge the crucial role of our teachers, whose instructions and guidelines acted as a foundation stone for this project.

Raneet Roy

Abhinav Kumar

Avishek Kumar Bose

Abstract

Breast cancer detection using machine learning (ML) has shown promising results in recent years and has a bright future. ML algorithms can improve accuracy, enable personalized medicine, and integrate with other technologies to provide a more comprehensive view of a patient's health. They can also help identify patients at high risk of developing breast cancer before any symptoms occur, enabling earlier interventions and potentially preventing the development of cancer altogether. ML algorithms can reduce the need for expensive tests and procedures, making breast cancer detection and treatment more cost-effective.

Keywords: Breast Cancer, Algorithms, Machine Learning

Contents

CHAPTER 1

Introduction

1.1 Overview

Breast cancer detection using machine learning has gained significant attention due to its potential to assist in early diagnosis and improve patient outcomes. Various machine learning algorithms can be employed for this purpose, and hyperparameter tuning techniques like Grid Search, Random Search, and Base Search can be applied to optimize the performance of these algorithms.

1. Machine Learning Algorithm:

Support Vector Machines (SVM):

Objective: SVM finds a hyperplane to separate data into classes.
Hyperplane: The line maximizing the margin between classes.
Support Vectors: Closest data points crucial for defining the hyperplane.
Margin: SVM maximizes the distance between hyperplane and data points.
Kernel Trick: Handles non-linear boundaries by mapping into higher-dimensional space.
C Parameter: Balances a smooth boundary and correct classification.
Gamma Parameter: Influences the radius for RBF kernel.
Optimization: Solves convex problem for optimal hyperplane.
Classification: Classifies new data based on their position relative to the hyperplane.

2. Data Preprocessing:

Feature Scaling:

- Normalize features to ensure that no single feature dominates the others.

Feature Selection:

- Identify and use relevant features to improve model efficiency.

Removing Null Values:

- Eliminate or impute missing data for a clean dataset before SVM model training.

3. Hyper-parameter Tuning Techniques:

Grid Search:

- Exhaustive search over a specified hyper-parameter grid.
- Define a grid of hyper-parameter values and evaluate the model's performance for each combination.

Random Search:

- Randomly samples from a defined hyper-parameter space.
- Allows for a more efficient search in high-dimensional spaces compared to Grid Search.

Base Search:

- Manually select hyper-parameter values based on prior knowledge or intuition.
- Less automated than Grid or Random Search but may be useful for small hyper-parameter spaces.

4. Evaluation Metrics:

Accuracy: - Proportion of correctly classified instances.

Precision, Recall, F1-Score: - Useful for imbalanced datasets, especially when dealing with medical diagnoses.

5. Implementation Steps:

A. Data Collection and Preprocessing:

- Collect and clean the dataset(breast-cancer-wisconsin.data).

- Handle missing values and outliers.

B. Feature Engineering:

- Extract relevant features or create new ones.

C. Model Selection:

- Choose appropriate algorithms based on the characteristics of the dataset.

D. Hyper-parameter Tuning:

- Apply Grid Search, Random Search, or Base Search to find the optimal hyper parameter values.

E. Model Training:

- Train the model on the training dataset using the optimized hyperparameters.

F. Evaluation:

- Assess the model's performance on a separate test dataset.

- Use appropriate evaluation metrics.

7. Challenges and Considerations:

Imbalanced Datasets:

- Address class imbalance to prevent biased models.

Computational Resources:

- Consider the computational cost associated with hyper-parameter tuning.

1.2 Significance and Applications of the project

Significance:

-Early Detection:

The project contributes to early detection of breast cancer, critical factor for successful treatment and improved patient outcomes.

-Reduced False Positives/Negatives:

Enhancing the accuracy of detection minimizes the chances of false positives (misdiagnosis) and false negatives (missed diagnoses), reducing unnecessary stress and ensuring timely intervention.

Application:

- Medical Diagnosis:

The primary application is in the medical field, where the model can assist healthcare professionals in identifying potential cases of breast cancer at an early stage.

- Patient Care:

Early detection allows for more effective treatment planning, potentially leading to better survival rates and improved quality of life for patients.

- Resource Optimization:

Efficient diagnosis helps in optimizing healthcare resources by focusing on cases that require attention, thus improving overall healthcare system efficiency.

1.3 Adding list

Software required:

(a) Google Colab

(b) Python framework

1.4 Adding tables

Table 1.1: *Accuracy,Precision,Recall,F1-Score for Grid Search Hyper-Parameter*

Hyper-Parameter	Accuracy	Precision	Recall	F1-Score
Grid-Search	0.95	0.93	0,94	0.93
Random-Search	0.96	0.93	0.95	0.94
Base-Search	0.96	0.92	0.97	0.94

CHAPTER 2

Project

2.1 IMPORT DATASET

```
import pandas as pd
df = pd.read_csv("breast-cancer-wisconsin.data",na_values = '?')
df.columns=["Sample code number","Clump Thickness","Uniformity of Cell
Size","Uniformity of Cell Shape","Marginal Adhesion","Single Epithelial
Cell Size","Bare Nuclei","Bland Chromatin","Normal
Nucleoli","Mitoses","Class"]
df
```

2.2 DATA ANALYSIS AND DATA PRE-PROCESSING

```
[ ] df.shape

    (698, 11)

[ ] df.size

    7678

[ ] df.info()

    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 698 entries, 0 to 697
    Data columns (total 11 columns):
     #   Column                       Non-Null Count  Dtype
    ---  ------                       --------------  -----
     0   Sample code number           698 non-null    int64
     1   Clump Thickness              698 non-null    int64
     2   Uniformity of Cell Size      698 non-null    int64
     3   Uniformity of Cell Shape     698 non-null    int64
     4   Marginal Adhesion            698 non-null    int64
     5   Single Epithelial Cell Size  698 non-null    int64
     6   Bare Nuclei                  682 non-null    float64
     7   Bland Chromatin              698 non-null    int64
     8   Normal Nucleoli              698 non-null    int64
     9   Mitoses                      698 non-null    int64
     10  Class                        698 non-null    int64
    dtypes: float64(1), int64(10)
    memory usage: 60.1 KB
```

```
# types of data
df.dtypes

Sample code number               int64
Clump Thickness                  int64
Uniformity of Cell Size          int64
Uniformity of Cell Shape         int64
Marginal Adhesion                int64
Single Epithelial Cell Size      int64
Bare Nuclei                    float64
Bland Chromatin                  int64
Normal Nucleoli                  int64
Mitoses                          int64
Class                            int64
dtype: object
```

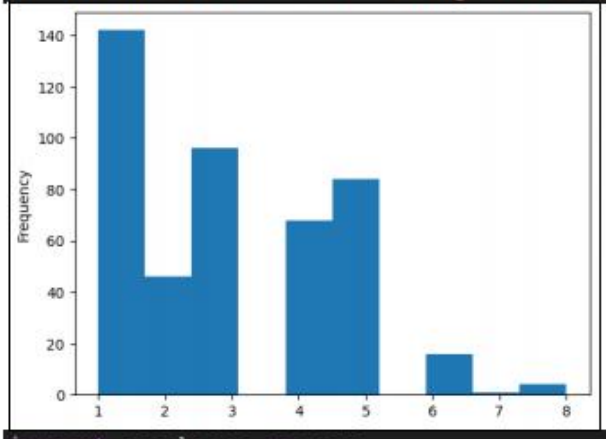
```
# checking for NaN in the df
df.isnull().sum().sort_values(ascending=False)

Bare Nuclei                    16
Sample code number              0
Clump Thickness                 0
Uniformity of Cell Size         0
Uniformity of Cell Shape        0
Marginal Adhesion               0
Single Epithelial Cell Size     0
Bland Chromatin                 0
Normal Nucleoli                 0
Mitoses                         0
Class                           0
dtype: int64
```

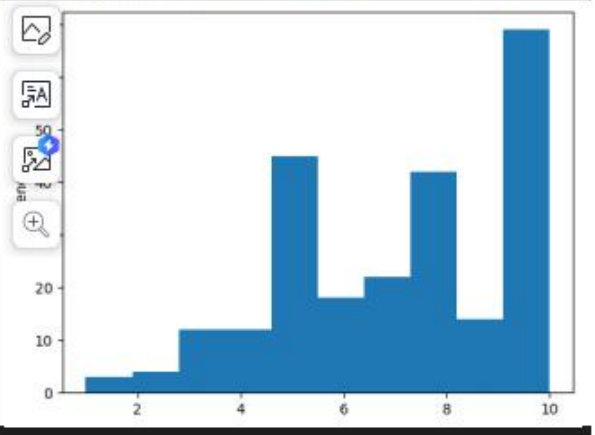
```
 #show NaN values
df[df.isna().any(axis=1)]
```

```
# auto insert mode values in place of NaN values
df2=df.fillna(df['Bare Nuclei'].mode()[0])

# again checking for NaN in the df after auto inserting
df2.isna().sum().sort_values(ascending=False)

# count distinct values in target collumn
df2['Class'].value_counts()
```

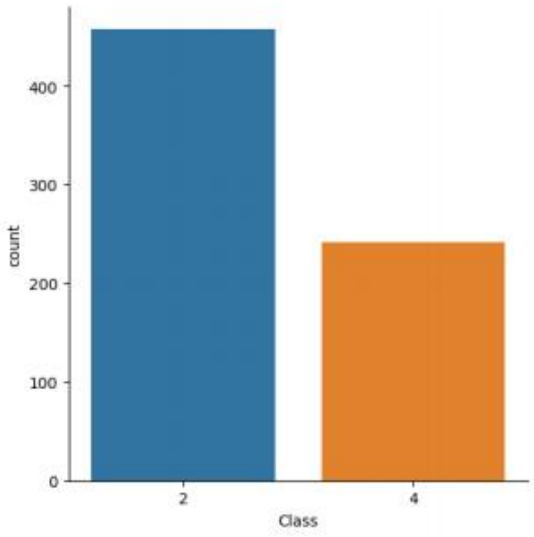
2.3 GRAPHS USING MATPLOTLIB AND SEABORN

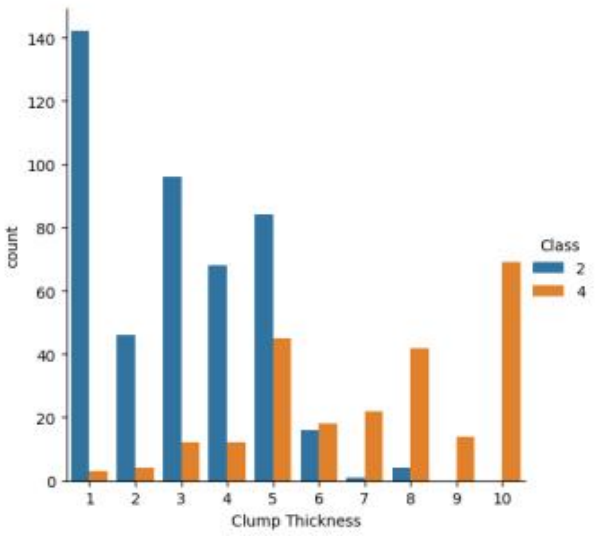
```
import matplotlib.pyplot as plt
df2[df2['Class']==2]['Clump Thickness'].plot(kind='hist')
df2[df2['Class']==4]['Clump Thickness'].plot(kind='hist')
```






```
import seaborn as sns
sns.catplot(x='Class',data=df2,kind='count')
sns.catplot(x='Clump Thickness',data=df2,kind='count',hue='Class')
```





2.4 SPLIT FEATURE AND TARGET COLUMNS

```
# features/columns
x=df2.iloc[:,1:10]
x
# target or labels
y=df2.iloc[:,-1]
y= y.map({2: 'B', 4: 'M'})
y
```

2.5 SPLIT DATASET IN THE RATIO TRAINING=70% TESTING =30%

```
# split data in 70% for training and 30% for testing
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=2)
#from sklearn.preprocessing import StandardScaler
#sc = StandardScaler()
#x_train = sc.fit_transform(x_train)
#x_test = sc.transform(x_test)
```

2.6 SVM WITHOUT TUNING

```
from sklearn import metrics
from sklearn.svm import SVC
clf=SVC(kernel='linear') #model creation
clf.fit(x_train,y_train) #training the model
y_pred3=clf.predict(x_test) #predicting the data using the model
print('Accuracy        :',metrics.accuracy_score(y_test, y_pred3)) #finding the accuracy
print('Precision Score :',metrics.precision_score(y_test, y_pred3,pos_label='M'))
print('Recall Score    :',metrics.recall_score(y_test, y_pred3,pos_label='M'))
print('F1 Score        :',metrics.f1_score(y_test, y_pred3,pos_label='M'))
```

```
Accuracy        : 0.9571428571428572
Precision Score : 0.9315068493150684
Recall Score    : 0.9444444444444444
F1 Score        : 0.9379310344827586
```

2.7 OPTIMIZE SVM USING GRID SEARCH

```
Best Parameters:  {'C': 1, 'gamma': 0.1, 'kernel': 'linear'}
Accuracy on Test Set:  0.9571428571428572
Precision on Test Set:  0.9315068493150684
Recall on Test Set:  0.9444444444444444
F1 Score on Test Set:  0.9379310344827586
```

2.8 OPTIMIZE SVM USING RANDOM SEARCH

```
Accuracy: 0.9619047619047619
Precision: 0.9324324324324325
Recall: 0.9583333333333334
F1 Score: 0.9452054794520548
```

2.9 OPTIMIZE SVM USING BASE SEARCH

```
Accuracy on Test Set:  0.9619047619047619
Precision on Test Set:  0.9210526315789473
Recall on Test Set:  0.9722222222222222
F1 Score on Test Set:  0.9459459459459458
```

Bibliography

1. U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control

2. Pritom, A.I., Munshi, M.A.R., Sabab, S.A., Shihab, S.: Predicting breast cancer recurrence using effective classification and feature selection technique. In: 19th International Conference on Computer and Information Technology (ICCIT), pp. 310–314. IEEE (2016)

3. Ojha U., Goel, S.: A study on prediction of breast cancer recurrence using data mining techniques. In: 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, IEEE, pp. 527–530, 2017

4. Muthu Rama Krishnan, Shuvo Banerjee, Chinmay Chakraborty, Chandan Chakraborty and Ajoy K. Ray, "Statistical analysis of mammographic features and its classification using support vector machine", *Expert Systems with Applications*, vol. 37, pp. 470-478, 2010.