

PROFESSIONAL TRAINING REPORT - II

entitled

BREAST CANCER COIMBRA

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering with
specialization in Internet of Things

by

VALLURI ADITHYA SAI [41732017]

N.H.K ARJUNA SAI [41732010]



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

CATEGORY -1 UNIVERSITY BY UGC

**Accredited with Grade "A++" by NAAC | 12B Status by UGC | Approved by AICTE
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119**

MAY 2024



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

CATEGORY-I UNIVERSITY BY UGC

Accredited "A++" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Professional Training Report is the bonafide work of **Mr./Ms. VALLURI ADITHYA SAI (41732017)** who carried out the project entitled "**BREAST CANCER COIMBRA**" under my supervision from January 2024 to May 2024.

Internal Guide

Mrs. K. Abhinaya, M.E.,(Ph.D)

Head of the Department

Dr. S. VIGNESHWARI, M.E., Ph.D.,

Submitted for Viva voce Examination held on 06/05/24

Internal Examiner

External Examiner

DECLARATION

I, VALLURI ADITHYA SAI (41732017), hereby declare that the Professional Training Report-II entitled "BREAST CANCER COIMBRA" done by me under the guidance of Mrs. K. Abhinaya M.E,(Ph.D), is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering with specialization in Internet of Things.

DATE: 06/05/2024

PLACE: Chennai

Valluri Adithya Sai
SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph.D., Dean, School of Computing, Dr. S. Vigneshwari M.E., Ph.D., Head of the Department of Computer Science and Engineering** for providing me necessary support during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Internal Guide < **Mrs.K.Abhinaya M.E,(Ph.D)**> for his/her valuable guidance, suggestions and constant encouragement which paved way for the successful completion of my phase-2 Professional Training.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

COURSE CERTIFICATE



ABSTRACT

Breast cancer is the most common cancer in women in developed countries, and 12% of breast cancer occurs in women 20-34 years. Survival from breast cancer has significantly improved, and the potential late effects of treatment and the impact on quality of life have become increasingly important. Young women constitute a minority of breast cancer patients, but commonly have distinct concerns and issues compared with older women, including queries regarding fertility, contraception and pregnancy.

Further, they are more likely than older women to have questions regarding potential side effects of therapy and risk of relapse or a new primary. In addition, many will have symptoms associated with treatment and they present a management challenge. Reproductive medicine specialists and gynaecologists commonly see these women either shortly after initial diagnosis or following adjuvant therapy and should be aware of current management of breast cancer, the options for women at increased genetic risk, the prognosis of patients with early stage breast cancer and how adjuvant systemic treatments may impact reproductive function.

There are many Algorithms in the field of Machine Learning (ML), we used Decision Tree Algorithm in classification to find accuracy of the data .In this study we find the last variable by the Decision tree classification and find the accuracy by undergoing some other methods in machine learning . Finally, we achieved the maximum performance score of our model i.e., Decision Tree classification.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	vi
	LIST OF FIGURES	viii
1	INTRODUCTION 1.1 Overview	1
2	LITERATURE SURVEY 2.1 survey	3
3	REQUIREMENTS ANALYSIS 3.1 Objective 3.2 3.2.1 Hardware Requirements 3.2.2 Software Requirements	4 5
4	DESIGN DESCRIPTION OF PROPOSED PRODUCT 4.1.1 Ideation Map/Architecture Diagram 4.1 4.1.2 Various stages 4.1.3 Internal or Component design structure 4.1.4 working principles 4.2 4.2.1 Novelty of the Project	6 10 12 14 15 17
5	RESULTS AND DISCUSSION	18
6	SUMMARY AND CONCLUSIONS	22
	REFERENCES	26
	APPENDIX A. Research Paper B. Source Code	27 33

LIST OF FIGURES

Figure No.	Figure Name	Page No.
1	FIG:1	8
2	FIG:2	9
3	FIG:3	10
4	FIG:4	15
5	FIG:5	19
6	FIG:6	20
7	FIG:7	20
8	FIG:8	21
9	FIG:9	23
10	FIG:10	24
11	FIG:11	25

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

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.

The older a woman is when she has her first child, the greater her risk of breast cancer.

Also at higher risk are:

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

BREAST CANCER COIMBRA DATASET

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. A few of the images can be found at

Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes

CHAPTER 2

LITERATURE REVIEW

2.1 SURVEY

Breast cancer was diagnosed in 2.3 million women, and around 685,000 deaths from breast cancer were recorded globally in 2020, making it the most common cancer. Early and accurate detection of breast cancer plays a critical role in improving the prognosis and bringing the patient survival rate to 50%. Deep learning-based computer-aided diagnosis (CAD) has achieved remarkable performance in early breast cancer diagnosis. This review focuses on literature considering deep learning architecture for breast cancer diagnosis. Therefore, this study anchors a well systematic and analytical review from six aspects: the model architecture of breast cancer diagnosis, datasets and image pre-processing, the manner of breast-cancer imaging, performance measurements, and research directions.

Breast Cancer Detection Methods

1. Digital Mammography (DM):

- X-ray imaging of breast tissue.
- Advantages: Widely available, relatively low cost.
- Disadvantages: Ionizing radiation exposure.

2. Infrared Thermography:

- Measures temperature variations in breast tissue.
- Advantages: Non-invasive, no radiation exposure.
- Disadvantages: Limited sensitivity.

3. Magnetic Resonance Imaging (MRI):

- Provides detailed images using magnetic fields and radio waves.
- Advantages: High resolution.
- Disadvantages: Expensive, time-consuming.

CHAPTER 3

REQUIREMENTS ANALYSIS

3.1 OBJECTIVE OF THE PROJECT

The objective of a project focusing on machine learning for breast cancer detection typically revolves around developing and/or evaluating algorithms and models that can accurately detect breast cancer from medical imaging data. Here are some specific objectives commonly associated with such projects:

1. **Developing Machine Learning Models:** The primary objective is often to develop machine learning models that can analyze medical images, such as mammograms, ultrasound images, or MRI scans, to detect the presence of breast cancer.
2. **Improving Accuracy:** Another objective is to improve the accuracy of breast cancer detection compared to traditional methods. This involves designing algorithms that can detect cancerous lesions with high sensitivity and specificity.
3. **Early Detection:** Early detection of breast cancer is crucial for improving patient outcomes. Projects may aim to develop models capable of identifying cancer at an early stage when treatment is most effective.
4. **Reducing False Positives:** Minimizing false positives is essential to avoid unnecessary anxiety and medical interventions for patients. The project may focus on reducing false positives while maintaining high sensitivity.

These objectives guide the design, implementation, and evaluation of machine learning algorithms and models for breast cancer detection projects, aiming to address critical challenges in medical imaging analysis and contribute to advancements in healthcare.

3.2 REQUIREMENTS

- ✚ Dataset
- ✚ Jupyter Notebook
- ✚ Programming skills-Python
- ✚ Access to a laptop or computer with internet connectivity

Quantitative Attributes:

- Age (years)
- BMI (kg/m²)
- Glucose (mg/dL)
- Insulin (μU/mL)
- HOMA

3.2.1 HARDWARE REQUIREMENTS

Hardware Requirements

- Any working Laptop
- Graphic processing unit
- Any Operating systems

3.2.2 SOFTWARE REQUIREMENTS

JUPYTER NOTEBOOK

DATASET

GOOGLE COLAB

Jupyter is a free, open-source, interactive web tool known as a computational notebook, which researchers can use to combine software code, computational output, explanatory text and multimedia resources in a single document.

CHAPTER 4

DESIGN DESCRIPTION OF PROPOSED PROJECT

4.1 PROPOSED METHODOLOGY

ANALYZING THE DATASET

- ✚ The first step in a data analytics or visualization project is to analyse the data being used.
- ✚ That is, the data must be thoroughly understood before working on it. ✚ Data analysis done by ✚ we must know the shape of the dataset being used. That is, getting to know the number of rows and columns in the entire dataset. In this case, it is 26706 rows and 27 columns as it was mentioned before.

DATA CLEANING

- ❖ Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.
- ❖ When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.
- ❖ We can either clean the data in the excel sheet before uploading it into the jupyter notebook, or we can clean the data directly in the notebook itself.
- ❖ There are many data cleaning techniques that we can use such as, removing missing and duplicate values, removing unnecessary columns, fill empty values.

MISSING VALUES

- Missing Data can occur when no information is provided for one or more items or for a whole unit.
-

- Missing Data is a very big problem in a real-life scenario. Missing Data can also refer to as NA (Not Available) values in pandas.
- In Pandas missing data is represented by two values: None, NaN (an acronym for Not a Number).
- Pandas treat None and NaN as essentially

interchangeable for indicating missing or null values. To facilitate this convention, there are several useful functions for detecting, removing, and replacing null values in Pandas Data Frame: `isnull()`, `dropna()`, `fillna()`, `replace()`.

After the cleaning of dataset, the dataset is now ready to be used for visualization, prediction purposes etc.

3.7 ALGORITHMS IN MACHINE LEARNING

In Machine Learning there are many classification algorithms to solve this problem. Examples of Classification algorithms are Logistic Regression, Knearest neighbours, decision tree, Random forest , etc.

3.7.1 LOGISTIC REGRESSION

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.
- Logistic regression uses the concept of predictive modelling as regression; therefore, it is called logistic regression, but is used to classify samples; Therefore, it falls under the classification algorithm.

Decision Tree

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcomes.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.

It is a graphical representation for getting all the possible solutions to a problem based on given conditions.

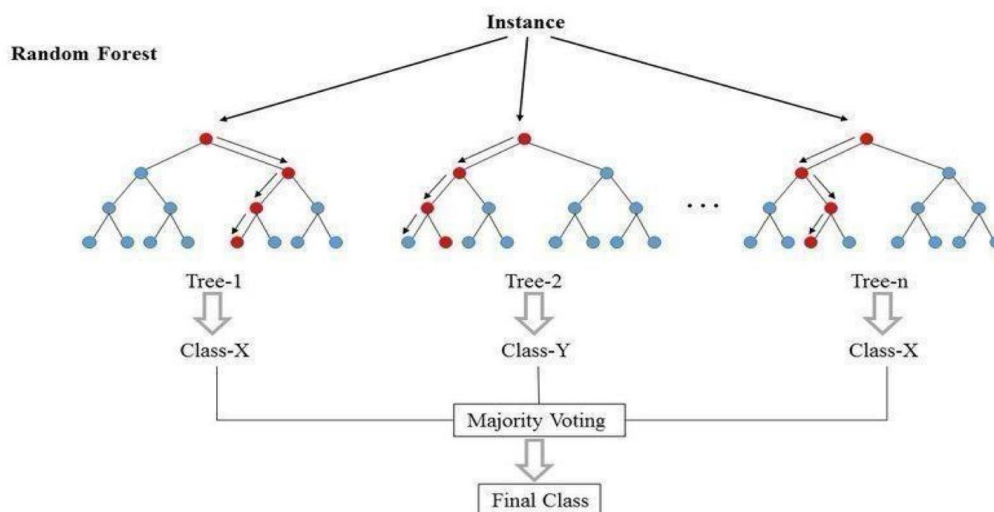
FIG:1



Random Forest

- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML.
- Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.
- Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
- Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

FIG:2

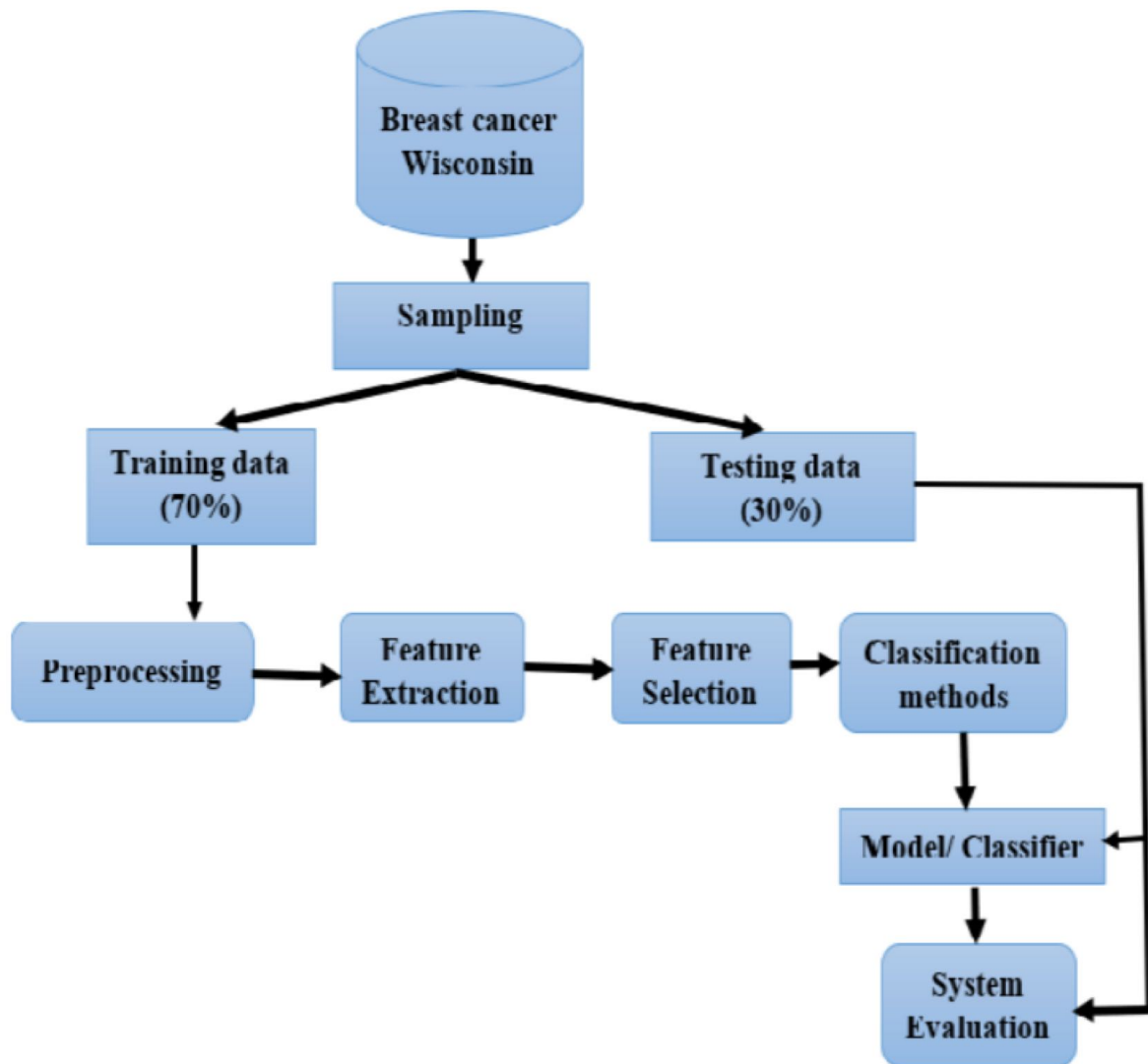


ALGORITHM USED

I have been asked to predict the breast cancer using machine learning by using both random forest and decision tree algorithms with the dataset used.

4.1.1 Ideation Map/System Architecture

FIG:3



Certainly! Here's an indentation map outlining the requirements for using the Random Forest algorithm in breast cancer detection:

1. Dataset

- Obtain a suitable dataset for training and testing the Random Forest model.

2. Data Preprocessing

- Preprocess the dataset to ensure cleanliness, standardization, and suitability for training.

- Steps may include resizing images, normalizing pixel intensities, and handling missing values.

3. Feature Extraction

- Extract relevant features from medical images for input into the Random Forest algorithm.

- Methods may include texture analysis, edge detection, or other image processing techniques.

4. Feature Selection

- Select informative features to improve the model's efficiency and effectiveness.

5. Model Training

- Train the Random Forest model using the preprocessed dataset and selected features.

- Split the dataset into training and validation sets.

- Optimize hyperparameters through techniques like cross-validation.

6. Evaluation Metrics

- Define evaluation metrics such as accuracy, sensitivity, specificity, AUC-ROC, and precision-recall curve.

7. Cross-Validation

- Perform cross-validation to assess the model's robustness and generalization performance.

This indentation map provides a structured overview of the requirements for using the Random Forest algorithm in breast cancer detection, facilitating clear understanding and organization of the tasks involved in the project.

4.1.2 Various Stages

Sure, here's an outline of the various stages involved in breast cancer detection using the Random Forest algorithm, with the execution done in Visual Studio Code (VS Code):

1. Data Collection and Preprocessing:

- Gather a dataset consisting of medical images of breast tissue (such as mammograms).
- Preprocess the images to ensure uniformity and quality. This may involve resizing, normalization, and noise reduction.

2. Feature Extraction:

- Extract relevant features from the preprocessed images. These features could include texture, shape, and intensity characteristics of the breast tissue.

3. Data Labeling:

- Label the images as either benign or malignant based on expert diagnosis or histopathological findings.

4. Model Development in VS Code:

- Set up a Python environment in VS Code for machine learning development.
- Write Python scripts to implement the Random Forest algorithm using libraries like scikit-learn.

5. Training the Random Forest Model:

- Split the dataset into training and testing sets.
- Train the Random Forest model using the training data. Tune hyperparameters as needed.

6. Model Evaluation:

- Evaluate the trained model's performance using metrics such as accuracy, sensitivity, specificity, and ROC curve.
- Use cross-validation techniques to ensure robustness and generalization of the model.

7. Interpretation and Visualization:

- Interpret the Random Forest model's results, including feature importance rankings.
- Visualize key metrics and performance characteristics using plots and graphs.

8. Deployment and Integration:

- Integrate the trained model into a user-friendly application or system.
- Utilize VS Code for deployment tasks such as version control, debugging, and testing.

By following these stages and utilizing VS Code for development and execution, you can effectively implement a breast cancer detection system using the Random Forest algorithm, leveraging machine learning techniques to aid in medical diagnosis.

4.1.3 Internal or Component design structure

1. Data Management Component:

- Responsible for handling the dataset, including data loading, preprocessing, and augmentation if necessary.

2. Feature Extraction Component:

- Extracts relevant features from the preprocessed medical images

3. Model Training Component:

- Trains the Random Forest model using the extracted features and corresponding labels

4. Model Evaluation Component:

- Evaluates the trained model's performance using various metrics.

5. Integration and Deployment Component:

- Integrates the trained model into a deployable system for real-world usage.

6. Validation and Testing Component:

- Validates the deployed system using independent datasets or real-world clinical data.

7. Documentation and Reporting Component:

- Documents the design, implementation, and evaluation of the system.

By designing the internal components in a modular and organized manner, you can facilitate the development, testing, and maintenance of the breast cancer detection system, ensuring scalability, reliability, and ease of use.

4.1.4 working principles

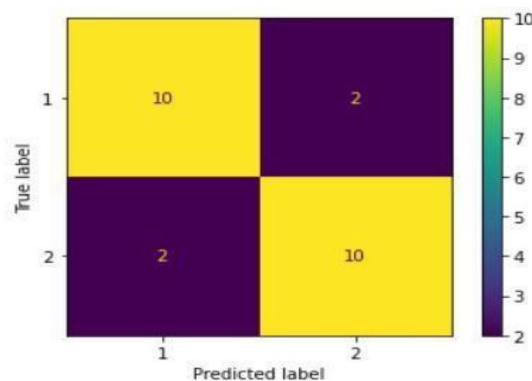
PERFORMANCE ANALYSIS

As you progress through the different aspects of data science, you will come across various evaluation metrics used to evaluate machine learning models. Machine learning models have to be evaluated in order to determine their effectiveness. You cannot run a machine learning model without evaluating it.

CONFUSION MATRIX

A confusion matrix is a tabular summary of the number of correct and incorrect predictions made by a classifier. It is used to measure the performance of a classification model.

FIG:4



PRECISION :

Precision explains how many correctly predicted values came out to be positive actually. Or simply it gives the number of correct outputs given by the model out of all the correctly predicted positive values by the model. It determines whether a model is reliable or not. It is useful for the conditions where false positive is a higher concern as compared to a false negative. For calculating the precision, the formula

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

RECALL:

$$\begin{aligned}\text{Recall} &= \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \\ &= \frac{\text{True Positive}}{\text{Total Actual Positive}}\end{aligned}$$

The ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples.

ACCURACY

One of the significant parameters in determining the accuracy of the classification problems, it explains how regularly the model predicts the correct outputs and can be measured as the ratio of the number of correct predictions made by the classifier over the total number of predictions made by the classifiers.

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

		Actual value	
		P	N
Prediction outcome	P	True Positive	False Positive
	N	False Negative	True Negative

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

4.2 FEATURES

Although machine learning has been transformative in some fields, machinelearning programs often fail to deliver expected results. Reasons for this are numerous: lack of (suitable) data, lack of access to the data, databias, privacy problems, badly chosen tasks and algorithms, wrong tools and people, lack of resources, and evaluation problems. Overfitting and Underfitting.

Our main objective in machine learning is to properly estimate the distribution and probability in training dataset so that we can have a generalized model that can predict the distribution and probability of the test dataset.

4.2.1 Novelty of the proposal

1. Integration of Machine Learning in Healthcare: The project bridges the gap between healthcare and technology by leveraging machine learning algorithms for medical diagnosis.

2. Application of Random Forest Algorithm: While machine learning techniques have been applied to medical imaging tasks, the specific use of the Random Forest algorithm for breast cancer detection
3. Customized Feature Extraction: The project involves extracting features from medical images tailored to breast cancer detection.
4. Adaptation to Clinical Practice: By integrating the developed system into clinical workflows, the project addresses real-world challenges in breast cancer diagnosis.

CHAPTER 5

RESULTS AND DISCUSSION

Although machine learning has been transformative in some fields, machine learning programs often fail to deliver expected results. Reasons for this are numerous: lack of (suitable) data, lack of access to the data, data bias, privacy problems, badly chosen tasks and algorithms, wrong tools and people, lack of resources, and evaluation problems. Overfitting and Underfitting.

Our main objective in machine learning is to properly estimate the distribution and probability in training dataset so that we can have a generalized model that can predict the distribution and probability of the test dataset.

While training models on a dataset, the most common problems people face are overfitting and underfitting. Overfitting is the main cause behind the poor performance of machine learning models. Overfitting refers to the scenario where a machine learning model can't generalize or fit well on unseen dataset. A clear sign of machine learning overfitting is if its error on the testing or validation dataset is much greater than the error on training dataset.

When the model neither learns from the training dataset nor generalizes well on the test dataset, it is termed as underfitting. This type of problem is not a headache as this can be very easily detected by the performance metrics.

Create a new Ipython Notebook and insert the below code to import the necessary modules. In case you get any error, do install the necessary packages .

RESULT

FIG:5

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [30]: from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
```

```
In [33]: from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
```

The given data set which is in the form of excel sheet is imported. Read the data using pandas into a data frame. To check the top 5 Rows of the dataset, use `df.head()`. You can specify the number of rows as an argument to this function in case you want to check different number of rows.

FIG:6

```
In [3]: df = pd.read_csv('dataR2.csv')
df.head()
```

```
Out[3]:
```

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
0	48	23.500000	70	2.707	0.467409	8.8071	9.702400	7.99585	417.114	1
1	83	20.690495	92	3.115	0.706897	8.8438	5.429285	4.06405	468.786	1
2	82	23.124670	91	4.498	1.009651	17.9393	22.432040	9.27715	554.697	1
3	68	21.367521	77	3.226	0.612725	9.8827	7.169560	12.76600	928.220	1
4	86	21.111111	92	3.549	0.805386	6.6994	4.819240	10.57635	773.920	1

DATA EXPLORATION

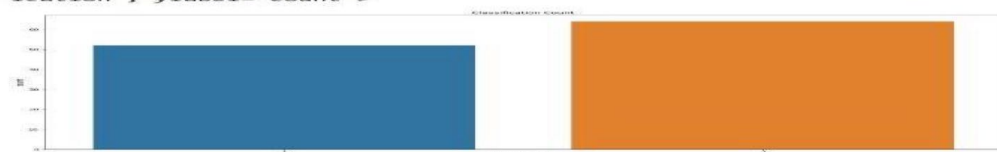
FIG:7

```
In [4]: df
Out[4]:
```

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
0	48	23.500000	70	2.707	0.467409	8.8071	9.702400	7.99585	417.114	1
1	83	20.690495	92	3.115	0.706897	8.8438	5.429285	4.06405	468.786	1
2	82	23.124670	91	4.498	1.009651	17.9393	22.432040	9.27715	554.697	1
3	68	21.367521	77	3.226	0.612725	9.8827	7.169560	12.76600	928.220	1
4	86	21.111111	92	3.549	0.805386	6.6994	4.819240	10.57635	773.920	1
...	...	...	...	...	...	...	...	...	...	...
111	45	26.850000	92	3.330	0.755688	54.6800	12.100000	10.96000	268.230	2
112	62	26.840000	100	4.530	1.117400	12.4500	21.420000	7.32000	330.160	2
113	65	32.050000	97	5.730	1.370399	51.4800	22.540000	10.33000	314.050	2
114	72	25.590000	82	2.820	0.570392	24.9600	33.750000	3.27000	392.460	2
115	66	27.160000	138	19.910	6.777364	90.2800	14.110000	4.35000	90.090	2

116 rows x 10 columns

```
Out[5]: <AxesSubplot:title={'center': 'Classification Count'}, xlabel='Classification', ylabel='count'>
```



```
Out[6]: <AxesSubplot:title={'center': 'Histogram of Age'}, xlabel='Age', ylabel='Count'>
```

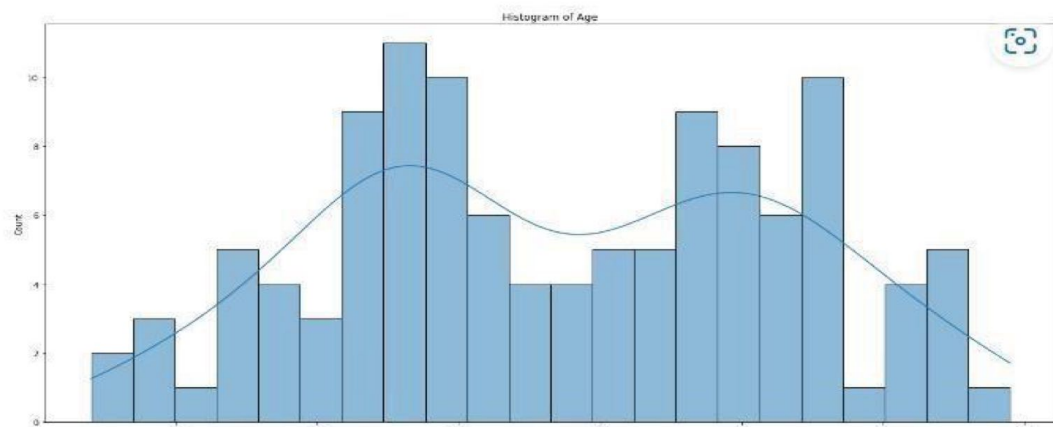
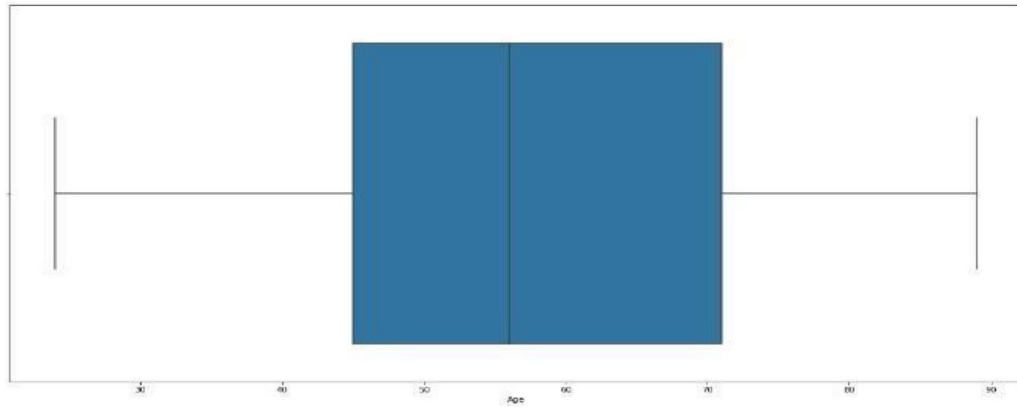
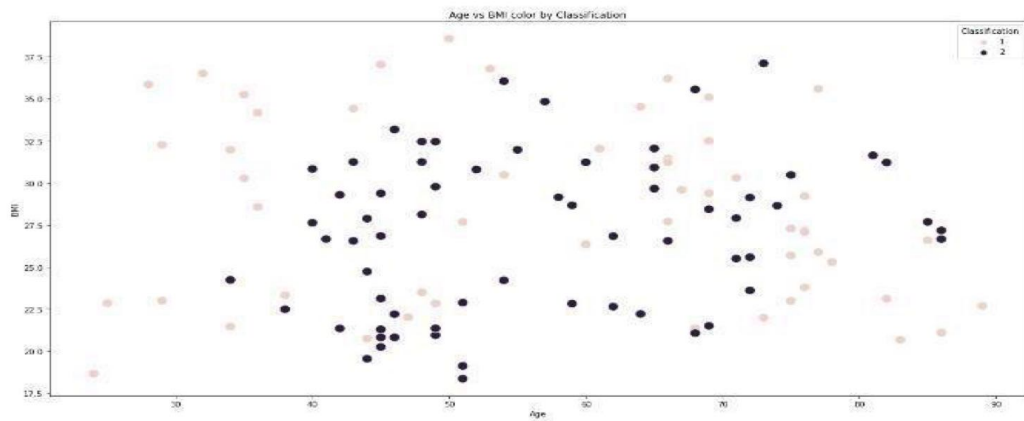


FIG:8

Out[7]: <AxesSubplot:xlabel='Age'>



Out[8]: <AxesSubplot:title={'center':'Age vs BMI color by Classification'}, xlabel='Age', ylabel='BMI'>



Out[9]: <AxesSubplot:title={'center':'Heatmap'}>



CHAPTER 5

CONCLUSION

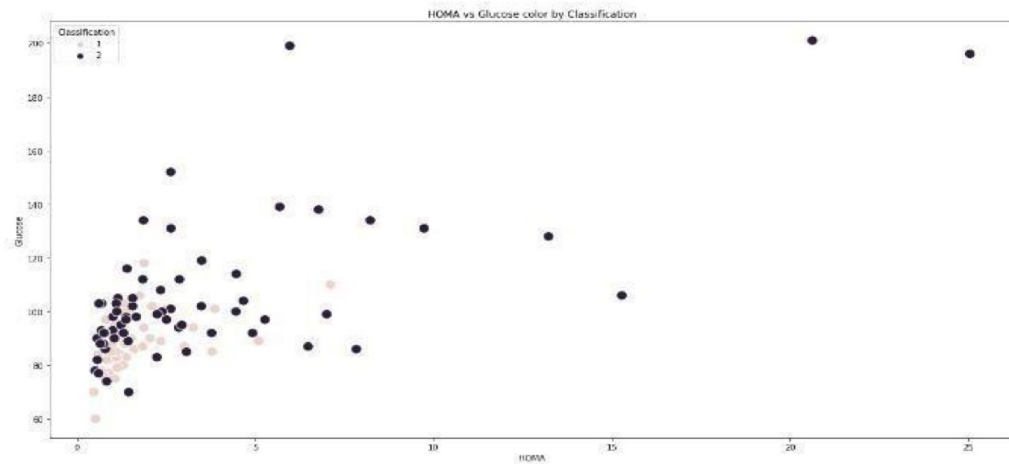
CONCLUSION

In the project in python, we learned to build a breast cancer tumour 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 very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable databases.

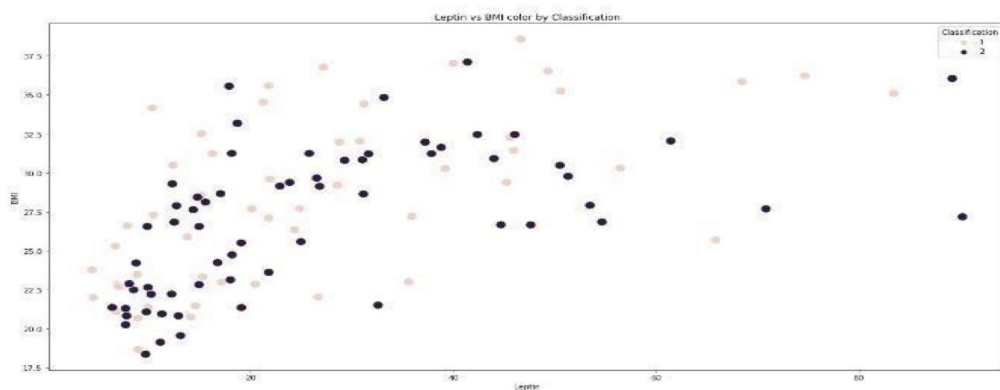
Breast Cancer represents one of the diseases that makes the highest number of deaths every year. At present, only a few accurate prognostic and predictive factors are used clinically for managing the patients with breast cancer. Here, by making use of Clustering with Level Set approach, high accuracy can be achieved in detection of affected cell shapes with exact marking on detected contours. The proposed system helps to enhance the performance of mammogram retrieval by selecting optimal features.

FIG:9

```
Out[10]: <AxesSubplot:title={'center':'HOMA vs Glucose color by Classification'}, xlabel='HOMA', ylabel='Glucose'>
```



```
Out[11]: <AxesSubplot:title={'center':'Leptin vs BMI color by Classification'}, xlabel='Leptin', ylabel='BMI'>
```



For the classification the correlation values are

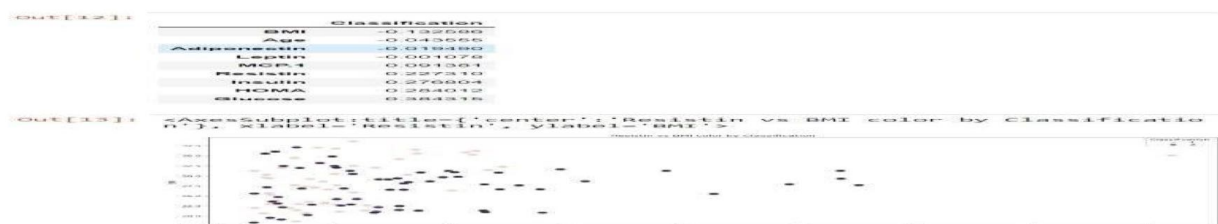
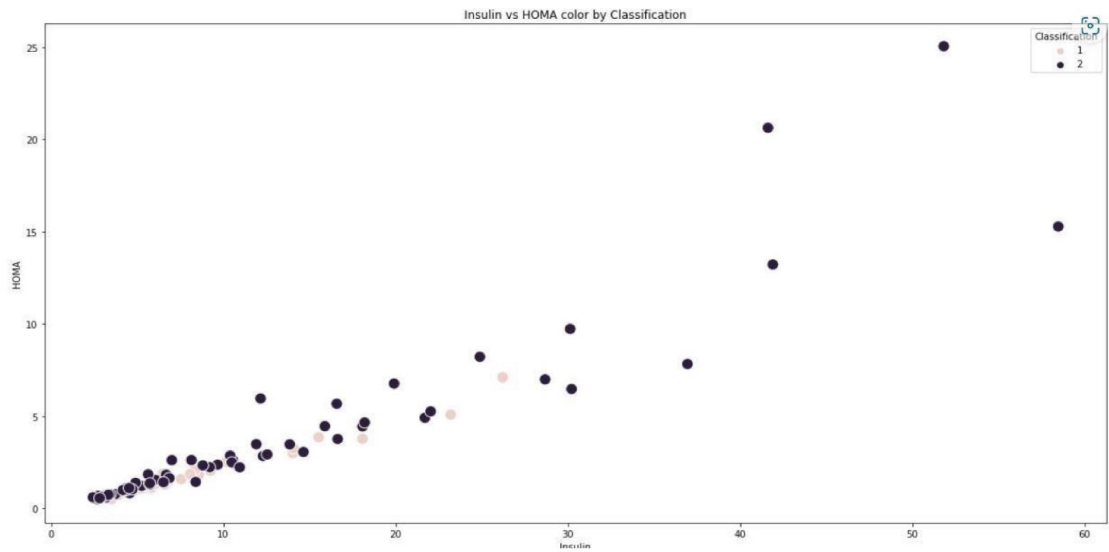
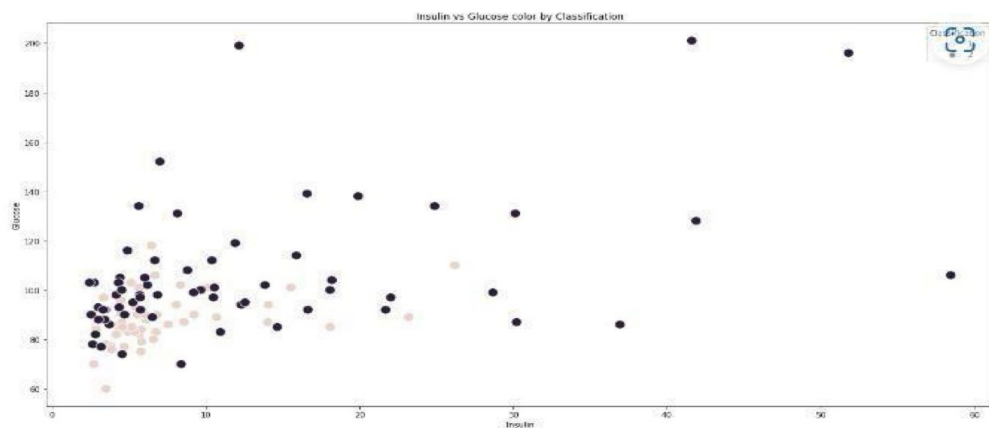


FIG:10

```
Out[14]: <AxesSubplot:title={'center':'Insulin vs HOMA color by Classification'}, xlabel='Insulin', ylabel='HOMA'>
```



```
Out[15]: <AxesSubplot:title={'center':'Insulin vs Glucose color by Classification'}, xlabel='Insulin', ylabel='Glucose'>
```



```
Out[16]: <AxesSubplot:title={'center':'Box Plot of the dataset'}}>
```

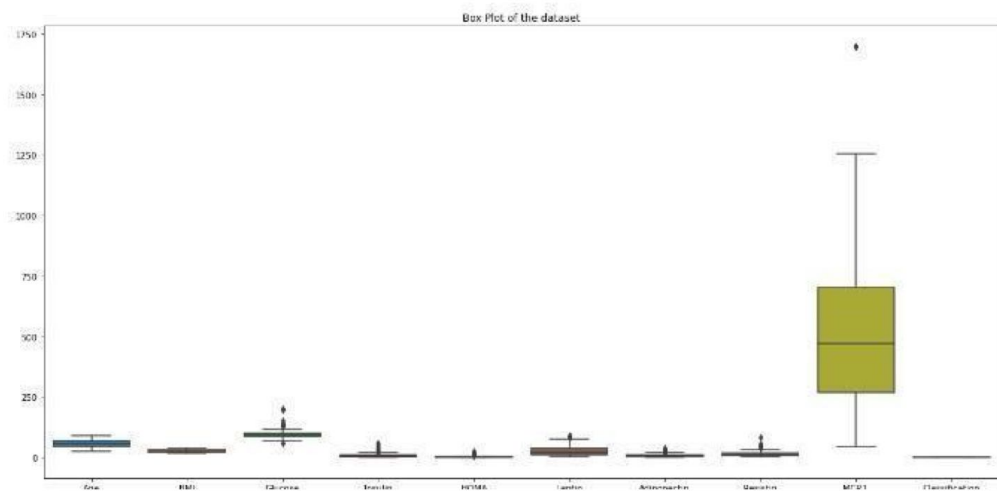
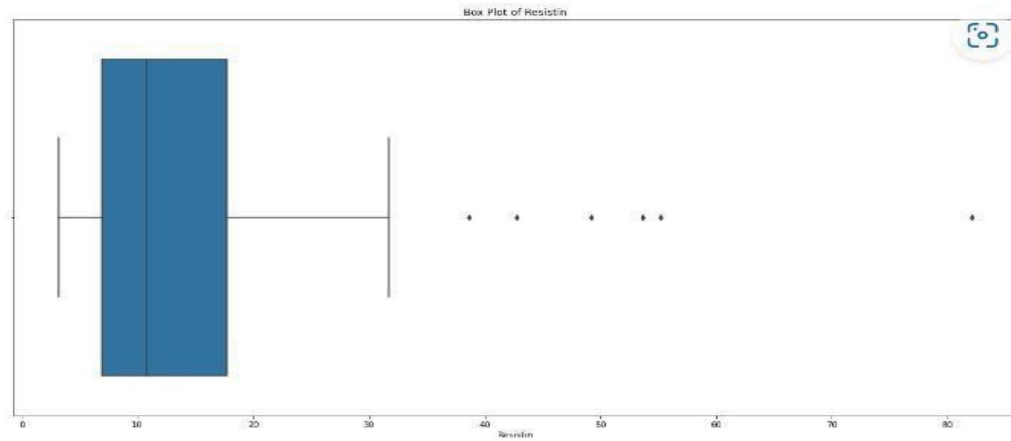


FIG:11

Out[20]: <AxesSubplot:title={'center': 'Box Plot of Resistin'}, xlabel='Resistin'>



Out[21]:

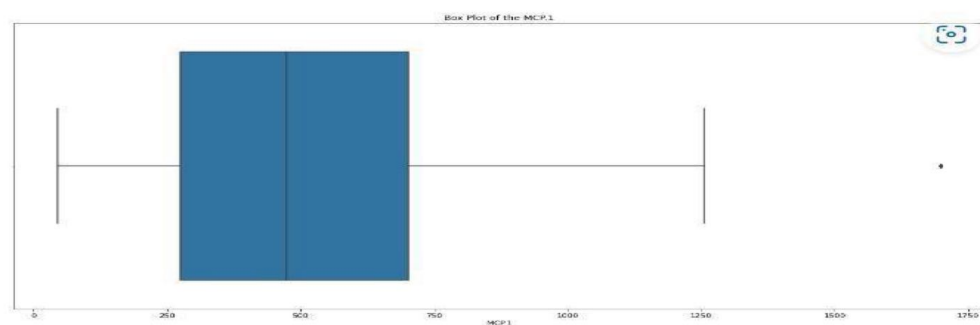
	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
37	69	35.092702	101	5.646	1.406607	83.4821	6.796985	82.10000	263
77	43	31.250000	103	4.328	1.099601	25.7816	12.718960	38.65310	775
83	71	25.510204	112	10.395	2.871792	19.0653	5.486100	42.74470	799
84	42	29.296875	98	4.172	1.008511	12.2617	6.695585	53.67170	1041
87	85	27.688778	196	51.814	25.050342	70.8824	7.901685	55.21530	1078
88	48	31.250000	199	12.162	5.969920	18.1314	4.104105	53.63080	1698
98	71	27.915519	104	18.200	4.668907	53.4997	1.656020	49.24184	256

MCP.1	Classification
263.499	1
775.322	2
799.898	2
1041.843	2
1078.359	2
1698.440	2
256.001	2

Out[22]:

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
37	69	35.092702	101	5.646	1.406607	83.4821	6.796985	82.1	263.499	1

Out[24]: <AxesSubplot:title={'center': 'Box Plot of the MCP.1'}, xlabel='MCP.1'>



REFERENCES

1. Zhang X, Shengli SU, Hongchao WA. Intelligent diagnosis model and method of palpation imaging breast cancer based on data mining. *Big Data Research* . 2019;5(1):2019005. doi: 10.11959/j.issn.2096-0271.2019005. [[CrossRef](#)] [[Google Scholar](#)]
2. Chen SI, Tseng HT, Hsieh CC. Evaluating the impact of soy compounds on breast cancer using the data mining approach. *Food & function* . 2020;11(5):4561–70. doi: 10.1039/C9FO00976K. [[PubMed](#)] [[CrossRef](#)] [[Google Scholar](#)]
3. Aavula R, Bhramaramba R, Ramula US. A Comprehensive Study on Data Mining Techniques used in Bioinformatics for Breast Cancer Prognosis. *Journal of Innovation in Computer Science and Engineering* . 2019;9(1):34–9. [[Google Scholar](#)]
4. Kaushik D, Kaur K. Application of Data Mining for high accuracy prediction of breast tissue biopsy results. and Wireless Communications (DIPDMWC); Moscow, Russia: IEEE; 2016. p. 40-
5. Mokhtar SA, Elsayad A. Predicting the severity of breast masses with data mining methods. *ArXiv preprint arXiv:1305* . 7057 2013
doi: 10.48550/ARXIV.1305.7057. [[CrossRef](#)] [[Google Scholar](#)]
6. Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology* . 2018;12(2):119–26.
doi: 10.1177/1748301818756225. [[CrossRef](#)] [[Google Scholar](#)]
7. Fan J, Wu Y, Yuan M, Page D, Liu J, Ong IM, Peissig P, Burnside E. Structure-leveraged methods in breast cancer risk prediction. *The Journal of Machine Learning Research* . 2016;17(1):2956–70. [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
8. Burnside ES, Liu J, Wu Y, Onitilo AA, McCarty CA, Page CD, et al. Comparing Mammography. doi: 10.1016/j.acra.2015.09.007. [[PMC Free Article](#)] [[PMC free article](#)] [[PubMed](#)] [[CrossRef](#)] [[Google Scholar](#)]
9. Stephens K. New Mammogram Measures of Breast Cancer Risk Could Revolutionize Screening. *AXIS Imaging News* . 2020 [[Google Scholar](#)]
10. Feld SI, Fan J, Yuan M, Wu Y, Woo KM, Alexandridis R, Burnside ES. Utility of Genetic Testing in Addition *AMIA Jt Summits Transl Sci Proc* . 2018;2017:81–90. [[PMC Free Article](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]

APPENDIX

1.RESEARCH PAPER

Problem Statement:

- 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 .
- 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.

ABSTRACT

Breast cancer is the most common cancer in women in developed countries, and 12% of breast cancer occurs in women 20-34 years. Survival from breast cancer has significantly improved, and the potential late effects of treatment and the impact on quality of life have become increasingly important. Young women constitute a minority of breast cancer patients, but commonly have distinct concerns and issues compared with older women, including queries regarding fertility, contraception and pregnancy.

Further, they are more likely than older women to have questions regarding potential side effects of therapy and risk of relapse or a new primary. In addition, many will have symptoms associated with treatment and they present a management challenge. Reproductive medicine specialists and gynaecologists commonly see these women either shortly after initial diagnosis or following adjuvant therapy and should be aware of current management of breast cancer, the options for women at increased genetic risk, the prognosis of patients with early stage breast cancer and how adjuvant systemic treatments may impact reproductive function.

There are many Algorithms in the field of Machine Learning (ML), we used Decision Tree Algorithm in classification to find accuracy of the data .In this study we find the last variable by the Decision tree classification and find the accuracy by undergoing some other methods in machine learning . Finally, we achieved the maximum performance score of our model i.e., Decision Tree classification.

Keywords

Keywords related to breast cancer detection using the Random Forest algorithm could include:

1. Breast cancer
2. Machine learning
3. Random Forest
4. Algorithm
5. Classification

INTRODUCTION

- Machine Learning(ML) is a branch of Artificial Intelligence (AI) & Computer Science, which focuses on the use of data and algorithms to imitate the way that humans learn and gradually improving accuracy.
- Improvement in classification accuracy have resulted from growing an ensemble of trees and letting them vote for the most popular class.
- Using random forest and decision algorithm we can get the accuracy of breast cancer.
- Decision Trees (DTs) are a non-parametric supervised learning method used for [classification](#) and [regression](#).

METHODOLOGY

There are many classifications algorithms in machine learning,for this project we selected random forest and decision to predict the result.

Step 1: Import the needed modules.

Step 2: Get your dataset and read it.

Step 3: Check your data, if there are any null or NaN values.

Step 4: Fill the null values in each column by calculating the mean each column.

Step 5: After filling the null values check your data if any columns are not needed and drop them.

Step 6: Once the data is clean now we are good to go with the prediction part.

Step7: Start prediction the data by considering Severity.

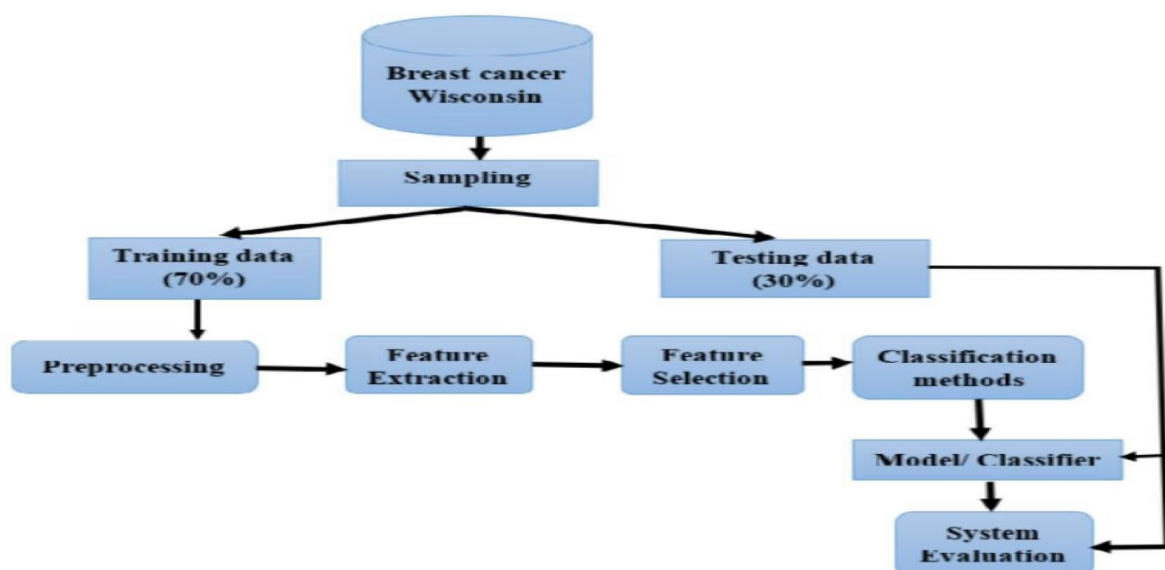
Step 8: Once predicting the data is completed start visualizing the data.

Step 9: Plot the graphs with the help of data and modules imported.

Step 10: Once visualizing the data is completed start analyzing the data.

Step 11: Generate the results.

SYSTEM ARCHITECTURE



DISCUSSION AND RESULTS

In the project in python, we learned to build a breast cancer tumour predictor on the Wisconsin dataset and created graphs and result 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 very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable databases.

Breast Cancer represents one of the diseases that makes highest number of deaths every year. At present, only few accurate prognostic and predictive factors are used clinically for managing the patients with breast cancer. Here, by making use of Clustering with Level Set approach, high accuracy can be achieved in detection of effected cell shapes with exact marking on detected contours. The proposed system helps to enhance the performance of mammogram retrieval by selecting optimal features.

Out[38]:

	Score
DecisionTreeClassifier	0.652174
RandomForestClassifier	0.739130
LogisticRegression	0.826087

Here we got accuracy_score 0.739130 for random forest and 0.652174 for the decision tree classifiers.

CONCLUSION

- A breast cancer tumor prediction is done using the Wisconsin dataset and created graphs are obtained with better accuracy.
- Selection of ML algorithms will lead to the development of prediction systems.
- These systems can assist in proper treatment methods for a doctor to diagnose breast cancer.
- This project will help the doctors and also the common people in deciding the line of treatment to be followed by extracting knowledge from such suitable databases.

REFERENCES

1. Zhang X, Shengli SU, Hongchao WA. Intelligent diagnosis model and method of palpation imaging breast cancer based on data mining. *Big Data Research* . 2019;5(1):2019005. doi: 10.11959/j.issn.2096-0271.2019005. [[CrossRef](#)] [[Google Scholar](#)]
2. Chen SI, Tseng HT, Hsieh CC. Evaluating the impact of soy compounds on breast cancer using the data mining approach. *Food & function* . 2020;11(5):4561–70. doi: 10.1039/C9FO00976K. [[PubMed](#)] [[CrossRef](#)] [[Google Scholar](#)]
3. Aavula R, Bhramaramba R, Ramula US. A Comprehensive Study on Data Mining Techniques used in Bioinformatics for Breast Cancer Prognosis. *Journal of Innovation in Computer Science and Engineering* . 2019;9(1):34–9. [[Google Scholar](#)]
4. Kaushik D, Kaur K. Application of Data Mining for high accuracy prediction of breast tissue biopsy results. and Wireless Communications (DIPDMWC); Moscow, Russia: IEEE; 2016. p. 40-
5. Mokhtar SA, Elsayad A. Predicting the severity of breast masses with data mining methods. *ArXiv preprint arXiv:1305* . 7057 2013 doi: 10.48550/ARXIV.1305.7057. [[CrossRef](#)] [[Google Scholar](#)]
6. Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology* . 2018;12(2):119–26. doi: 10.1177/1748301818756225. [[CrossRef](#)] [[Google Scholar](#)]

7. Fan J, Wu Y, Yuan M, Page D, Liu J, Ong IM, Peissig P, Burnside E. Structure-leveraged methods in breast cancer risk prediction. *The Journal of Machine Learning Research* . 2016;17(1):2956–70. [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
8. Burnside ES, Liu J, Wu Y, Onitilo AA, McCarty CA, Page CD, et al. Comparing Mammography.doi: 10.1016/j.acra.2015.09.007. [[PMC Free Article](#)] [[PMC free article](#)] [[PubMed](#)] [[CrossRef](#)] [[Google Scholar](#)]
9. Stephens K. New Mammogram Measures of Breast Cancer Risk Could Revolutionize Screening. *AXIS Imaging News* . 2020 [[Google Scholar](#)]
10. Feld SI, Fan J, Yuan M, Wu Y, Woo KM, Alexandridis R, Burnside ES. Utility of Genetic Testing in Addition *AMIA Jt Summits Transl Sci Proc* . 2018;2017:81–90. [[PMC Free Article](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]

2. SOURCE CODE

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os
for dirname, _, filenames in
os.walk('/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

Reading the dataset df =
pd.read_csv('dataR2.csv')
df.head()

Data Exploration
plt.figure(figsize=(20,10))
plt.title('Classification Count')
sns.countplot(data=df, x='Classification')

plt.figure(figsize=(20,10))
plt.title('Histogram of Age')
sns.histplot(data=df, x='Age', bins=22, kde=True)

plt.figure(figsize=(20,10))
sns.boxplot(data=df, x='Age')

plt.figure(figsize=(20,10))
```

```

plt.title('Age
vs BMI color
by
Classification'
)
sns.scatterplo
t(data=df,
x='Age',y='B
MI',
hue='Classific
ation',s=150)
plt.figure(figsize=(20,10))
plt.title('Heat
map')
sns.heatmap(
data=df.corr()
, annot=True)
plt.figure(figsize=(20,10))
plt.title('HOM
A vs Glucose
color by
Classification'
)
sns.scatterplo
t(data=df,
x='HOMA',y='
Glucose',
hue='Classific
ation',s=150)
plt.figure(figsize=(20,10))
plt.title('Leptin

```

```

vs BMI color
by
Classification'
)
sns.scatterplo
t(data=df,
x='Leptin',y='
BMI',
hue='Classific
ation',s=150)
df.corr()[['Clas
sification']].so
r
t_values('Clas
sification',
ascending=
True)[-1]
plt.figure(figsize=(20,10))
plt.title('Resist
in vs BMI
color by
Classification'
)
sns.scatterplo
t(data=df,
x='Resistin',y
='BMI',
hue='Classific
ation',s=150)
plt.figure(figsize=(20,10))
plt.title('Insuli
n vs HOMA

```

```

color by
Classification'
)
sns.scatterplo
t(data=df,
x='Insulin',y='
HOMA',
hue='Classific
ation',s=150)
plt.figure(figsize=(20,10))
plt.title('Insuli
n vs Glucose
color by
Classification'
)
sns.scatterplo
t(data=df,
x='Insulin',y='
Glucose',
hue='Classific
ation',s=150)
plt.figure(figsize=(20,10))
plt.title('Box
Plot of the
dataset')
sns.boxplot(d
ata=df)
plt.figure(figsize=(20,10))
plt.title('Box
Plot of the
dataset')

```

```

sns.boxplot(data=df.drop(['MCP.1', 'Classification'],
axis=1))
plt.figure(figsize=(20, 10))
plt.title('Box Plot of Glucose')
sns.boxplot(data=df,
x='Glucose')
df[(df['Glucose'] < 70) |
(df['Glucose'] >
120)]
plt.figure(figsize=(20, 10))
plt.title('Box Plot of Resistin')
sns.boxplot(data=df,
x='Resistin')
df[(df['Resistin'] > 35) |
(df['Resistin'] > 80)]
df = df.drop(df[(df['Resistin'] > 80)].index)
plt.figure(figsize=(20, 10))
plt.title('Box Plot of the MCP.1')
sns.boxplot(data=df,
x='MCP.1')
df[(df['MCP.1'] > 750)].sort_values('MCP.1')
df[(df['MCP.1'] > 1500)]
df = df.drop(df[(df['MCP.1'] > 1500)].index)
df

```

Modelling

```
X = df.drop('Classification', axis=1) y
= df['Classification']
from sklearn.model_selection import
train_test_split
from sklearn.preprocessing import
StandardScaler
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.2, random_state=42)
scaler = StandardScaler()
```

```
X_train = scaler.fit_transform(X_train)
```

```
X_test = scaler.transform(X_test)
```

```
models = {'RandomForestClassifier': RandomForestClassifier(),
          'DecisionTreeClassifier': DecisionTreeClassifier(), }
baseline_model_scores = fit_and_score(models, X_train,
                                       X_test, y_train, y_test)
baseline_model_scores
plt.figure(figsize=(20,10))
sns.barplot(data=baseline_model_scores.sort_values('Score').T)
plt.title('Baseline Model Precision Score')
plt.xticks(rotation=90);
```

Random Search CV

```
from sklearn.model_selection import
RandomizedSearchCV
def randomsearch_cv_scores(models,
                           params, X_train,
                           X_test, y_train, y_test):
    np.random.seed(42)
    model_rs_scores = {}
    model_rs_best_param = {}
```

```

for name, model in models.items(): rs_model =
    RandomizedSearchCV(model,
        param_distributions=params[name],
        verbose=0)

    rs_model.fit(X_train,y_train) model_rs_scores[name] =
    rs_model.score(X_test,y_test)

    model_rs_best_param[name] = rs_model.best_params_
return model_rs_scores,
model_rs_best_param

```

Model Evaluation

```

from sklearn.metrics import
classification_report,
    plot_confusion_matrix,plot_roc_curve from
    sklearn.model_selection import cross_val_score model =
    LogisticRegression(C=0.2, l1_ratio=0.25,
    penalty='elasticnet',solver='saga')
    model.fit(X_train,y_train) y_preds = model.predict(X_test)

```

Classification Report

```

print(classification_report(y_test,y_preds))

```

Confusion Matrix

```

plot_confusion_matrix(model, X_test, y_test)
Validation accuracy Mean Score: {cv_accuracy.mean()}'
    cv_precision = cross_val_score(model,X,y,cv=cv,

        scoring='precision')

    print(f'Cross Validation precision Scores: {cv_precision}')
    print(f'Cross

```

Validation precision Mean Score:

```

    {cv_precision.mean()}' cv_recall =
    cross_val_score(model,X,y,cv=cv,
        scoring='recall')

```

```

print(f'Cross Validation recall Scores: {cv_recall}') print(f'Cross

```

```

Validation recall Mean Score: {cv_recall.mean()})
cv_f1 = cross_val_score(model,X,y,cv=cv,

                        scoring='f1')

print(f'Cross Validation f1 Scores: {cv_f1}') print(f'Cross

Validation f1 Mean Score: {cv_f1.mean()})
cv_merics = pd.DataFrame({'Accuracy':
cv_accuracy.mean(),

                        'Precision': cv_precision.mean(),

                        'Recall': cv_recall.mean(),

                        'f1': cv_recall.mean()},index=[0])

return cv_merics cv_merics =
get_cv_score(model, X_train, y_train,
cv=5)
cv_merics

```