

# VISVESVARAYA TECHNOLOGICAL UNIVERSITY

“JnanaSangama”, Belgaum -590014, Karnataka.

## PROJECT WORK-4 REPORT

on

”Breast Cancer Detection Using Machine Learning”

Submitted by

**Kizhakel Sharat Prasad (1BM19CS074)**

**M Vamshi Krishna (1BM19CS080)**

Under the guidance of

**Dr.Selva Kumar S**

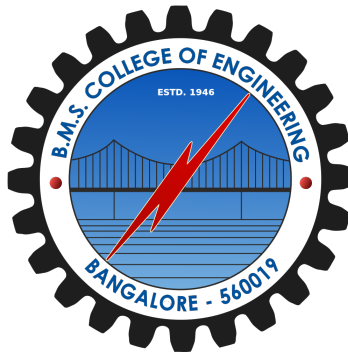
Assistant Professor,BMSCE

in partial fulfilment of the requirements for the degree of

**Bachelor of Engineering**

in

**COMPUTER SCIENCE AND ENGINEERING**



**B. M. S. COLLEGE OF ENGINEERING**

(An autonomous institution affiliated to VTU, Belagavi)

Bull Temple Road, Basavanagudi, Bengaluru - 560 019

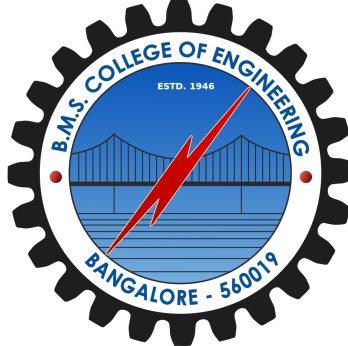
April-2022 to July-2022

# B. M. S. COLLEGE OF ENGINEERING

Bull Temple Road, Basavanagudi, Bengaluru - 560 019

(Affiliated To Visvesvaraya Technological University, Belgaum)

**Department of Computer science and Engineering**



## Certificate

This is to certify that the project work entitled “**Breast Cancer Detection Using Machine Learning**” carried out by **Kizhakel Sharat Prasad (1BM19CS074)** AND **M Vamshi Krishna (1BM19CS080)** who are bonafide students of B. M. S. College of Engineering. It is in partial fulfillment for the award of Bachelor of Engineering in Computer Science and Engineering of the Visveswararaja Technological University, Belgaum during the year 2021. The project report has been approved as it satisfies the academic requirements in respect of Project Work-4 (20CS6PWPW4) work prescribed for the said degree.

Signature of the HOD

Prof. Namratha M

Assistant Professor

BMSCE, Bengaluru

Signature of the HOD

Dr. Jyothi S Nayak

Professor Head, Dept. of CSE

BMSCE, Bengaluru

External Viva

Name of the Examiner

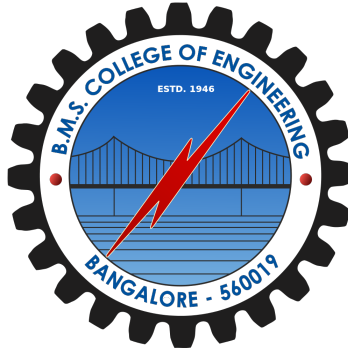
Signature with Date

1.

2.

## **B. M. S. COLLEGE OF ENGINEERING**

### **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**



### **DECLARATION**

We, Kizhakel Sharat Prasad (1BM19CS074) AND M Vamshi Krishna (1BM19CS080), students of 6th Semester, B.E, Department of Computer Science and Engineering, B. M. S. College of Engineering, Bangalore, hereby declare that, this Project Work4 entitled "Breast Cancer Detection Using Machine Learning" has been carried out by us under the guidance of Dr.Selva Kumar, Professor, Department of CSE, B. M. S. College of Engineering, Bangalore during the academic semester Apr-2022-Aug-2022

Kizhakel Sharat Prasad (1BM19CS074)

M Vamshi Krishna (1BM19CS080)

Place:

Date:

# Contents

List of figures . . . . .	ii
List of tables . . . . .	iii
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Scope of the Project . . . . .	2
1.3 Problem Statement . . . . .	3
<b>2 Literature Survey</b>	<b>4</b>
<b>3 Design</b>	<b>6</b>
3.1 High Level Design . . . . .	6
3.2 Detailed Design . . . . .	8
3.3 Sequence Design . . . . .	9
3.4 Usecase Design . . . . .	10
<b>4 Implementation</b>	<b>12</b>
4.1 Proposed methodology . . . . .	12
4.2 Algorithm used for implementation . . . . .	12
4.3 Tools and technologies used . . . . .	14
4.4 Screenshots . . . . .	15
4.5 Testing . . . . .	17
<b>5 Results and Discussion</b>	<b>18</b>
5.1 ROC . . . . .	20

# List of Figures

2.1	Comparison table . . . . .	5
3.1	High Level Design . . . . .	7
3.2	Detailed Design . . . . .	8
3.3	Sequence Design . . . . .	9
3.4	Usecase Design . . . . .	10
4.1	Logistic Regression Classification Report . . . . .	13
4.2	Decision Tree Classification Report . . . . .	13
4.3	Random Forest Classification Report . . . . .	14
5.1	Result After Dropping 120 Rows . . . . .	19
5.2	Result After Dropping 2 Columns . . . . .	19
5.3	Result After Dropping 15 Columns . . . . .	19
5.4	ROC Curve . . . . .	20

# List of Tables

# Chapter 1

## Introduction

Total number of women dying in 2021 is approximately 963,000, according to the World Health Organization (WHO). Still, the organization predicts that the number could reach 2.9 million globally. Breast cancer can occur in women and rarely in men. The ICMR (Indian Council of Medical Research) recently published a report which stated that in 2020 the total number of new cancer cases is expected to be about 17.3 lakhs. An Indian woman is diagnosed with breast cancer in every four minutes. Breast cancer is a disease that occurs but when a woman or a man is aware of this symptom, it immediately goes beyond its original stage. Breast cancer is a common and dangerous disease in women, cancer is the creation of abnormal cells that come into these cells genetically and mutated. Spreads throughout the body, leading to death in diagnosis and treatment. There are two types of breast cancer, Malignant and Benign. The first is classified as harmful has the ability to infect other organs and is cancerous, Benign is classified as non-cancerous. This disease infects the women's chest and specifically glands and milk ducts, the spread of breast cancer to other organs is frequent and could be through the bloodstream. Different techniques are used to capture breast cancer such as Ultrasound Sonography, Computerized Thermography, Biopsy (Histological images).

### 1.1 Motivation

According to the Centers for Disease Control and Prevention (CDC) Trusted Source, breast cancer is the most common cancer in women. Breast cancer survival rates vary widely supported by many factors. Two of the most important factors are the type of cancer women have and the stage of cancer at the time they receive a diagnosis. Breast cancer is cancer that develops in breast cells. Typically, the cancer forms in either the lobules or the ducts of the breast. Cancer also can occur within the adipose tissue or the fibrous connective tissue within your breast. The uncontrolled cancer cells often invade other healthy breast tissue and may visit the lymph nodes under the arms. Doctors say that breast cancer happened due to abnormal growth of cells in the breast and these cells spread in size like Meta Size from breast to lymph nodes or

the other parts of the body also. Hence it is necessary to detect and stop the growth of these unwanted cells as early as possible to avoid the next phase consequences.

Breast cancer is a widely occurring cancer in women worldwide and is related to high mortality. The objective of this review was to present several approaches to investigate the application of multiple algorithms based on machine learning (ML) approach for early breast cancer detection

## **1.2 Scope of the Project**

There are two types of breast cancer, Malignant and Benign. The first is classified as harmful has the ability to infect other organs and is cancerous, Benign is classified as non-cancerous. This disease infects the women's chest and specifically glands and milk ducts, the spread of breast cancer to other organs is frequent and could be through the bloodstream. A variety of techniques are used to capture breast cancer such as Ultrasound Sonography, Computerized Thermography, Biopsy. Machine learning and Data mining techniques are straightforward and effective ways to understand and predict data. Radiologist examines and analyses himself and then he / she decides the result after participating with other experts. This process takes time and the results depend on the knowledge and experience of the staff. In addition there is a dearth of experts in every field of the world. Thus the research community made a proposal for automatic A system called CAD (Computer-Aided Diagnosis) for better classification of tumours, accurate results and faster Implementation without the need for radiologists or specialists. Machine learning algorithms (MLs) are indicated as one Option of human vision and experience to make final decisions with high accuracy. 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.



## 1.3 Problem Statement

if the cancer diagnosis is benign or malignant based on several observations/features.

We extract out of the images some features, when we see features that mean some characteristics out of the image such as radius, for example the cells such as texture perimeter area smoothness and so on. And then we feed all these features into kind of our machine learning model. Some features to be considered are: radius (mean of distances from center to points on the perimeter)

texture (standard deviation of gray-scale values) - perimeter

area - smoothness (local variation in radius lengths)

compactness ( $\text{perimeter}^2/\text{area} - 1.0$ )

*concavity(severityofconcaveportionsofthecontour)*

*concavepoints(numberofconcaveportionsofthecontour)*

*symmetry*

*fractaldimension("coastlineapproximation" - 1)*

*Datasetsarelinearlyseparableusingall30inputfeatures*

**Number of Instances: 569**

**Class Distribution: 212 Malignant, 357 Benign**

**Target class: Malignant - Benign**

## Chapter 2

# Literature Survey

On the Wisconsin Breast Cancer datasets, two main algorithms, which are: NB(Naïve Bayes) KNN, since our target and challenge from breast cancer classification is to build classifiers that are precise and reliable. After an accurate comparison between our algorithms, we noticed that KNN achieved a higher efficiency of 97.51%, however, even NB has a good accuracy at 96.19%, if the dataset is larger, the KNN's time for running will increase. BCC aims to determine the suitable treatment, which can be aggressive or less aggressive, depending on the class of the cancer. To make a good prognostic, breast cancer classification needs nine characteristics which are:

1. Determine the layered structures (Clump Thickness).
2. Evaluate the sample size and its consistency (Uniformity of Cell Size).
3. Estimate the equality of cell shapes and identifies marginal variances, because cancer cells tend to vary in shape (Uniformity of Cell Shape).
4. Cancer cells spread all over the organ and normal cells are connected to each other (Marginal Adhesion).
5. Measure of the uniformity, enlarged epithelial cells are a sign of malignancy (Single Epithelial Cell Size).
6. In benign tumors nuclei is not surrounded by cytoplasm (Bare Nuclei).
7. Describes the nucleus texture, in benign cells it has a uniform shape. The chromatin tends to be coarser in tumors (Bland Chromatin).
8. In normal cells, the nucleolus is usually invisible and very small. In cancer cells, there are more than one nucleoli and it becomes much more prominent, (Normal Nucleoli).
9. Estimate of the number of mitosis that has taken place. The larger the value, the greater is the chance of malignancy (Mitoses).

Comparison Between Techniques		
Techniques	Accuracy Without Standard scale	Accuracy With Standard scale
SVM	57.89%	96.49%
KNN	93.85%	57.89%
Random Forest	97.36%	75.43%
Decision Tree	94.73%	75.43%
Naïve Bayes	94.73%	93.85%
Adaboost	94.73%	94.73%
XGboost	98.24%	98.24%

Figure 2.1: Comparison table

The objectives of the paper was to analyse the Wisconsin breast cancer dataset by visualizing and evaluating Machine Learning Predictions. With this research paper we can see that among Naïve Bayes, Support Vector Machine, Adaboost, Random Forest Classifier, KNN, Decision Tree, XGboost etc. They concluded that XGboost is the most accurate algorithm for best accurate result for detection of breast cancer with the efficiency of 98.24%. However, it is required that before running the algorithm, the dataset must be pre-processed. Their Methods Includes Supervised Learning Algorithms and Classification Techniques like Support Vector Classifier (SVM), Random Forest, Naïve Bayes, Decision Tree, KNN, Adaboost and XGboost. Dataset contains features which highly vary in units and magnitudes. So, it is required to bring all features to the same level of magnitudes. They achieved that by using Standard Scaling in SKLearn[1].

This paper proposes a hybrid model combined of several Machine Learning (ML) algorithms including Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Decision Tree (DT) for effective breast cancer detection. This study also discusses the datasets used for breast cancer detection and diagnosis. The proposed model can be used with different data types such as image, blood, etc. The findings of these researchers suggest that SVM is the most popular method used for cancer detection applications. SVM was used either alone or combined with another method to improve the performance. The maximum achieved accuracy of SVM (single or hybrid) was 99.8% that can be improved to 100%. It was observed from the work of who used optional ANN on MRI resulted in 100% accuracy in detecting breast cancer. This method can be applied and tested on another dataset like mammogram and ultrasound to check the performance of different data types[2].

## Chapter 3

# Design

### 3.1 High Level Design

We can define the machine learning workflow in 5 stages.

1. Gathering data
2. Data pre-processing
3. Researching the model that will be best for the type of data
4. Training and testing the model
5. Evaluation

Data pre-processing is one of the most important steps in machine learning. It is the most important step that helps in building machine learning models more accurately. In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time for data pre-processing and 20% time to actually perform the analysis. The data set can be collected from various sources such as a file, database, sensor and many other such sources but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data.

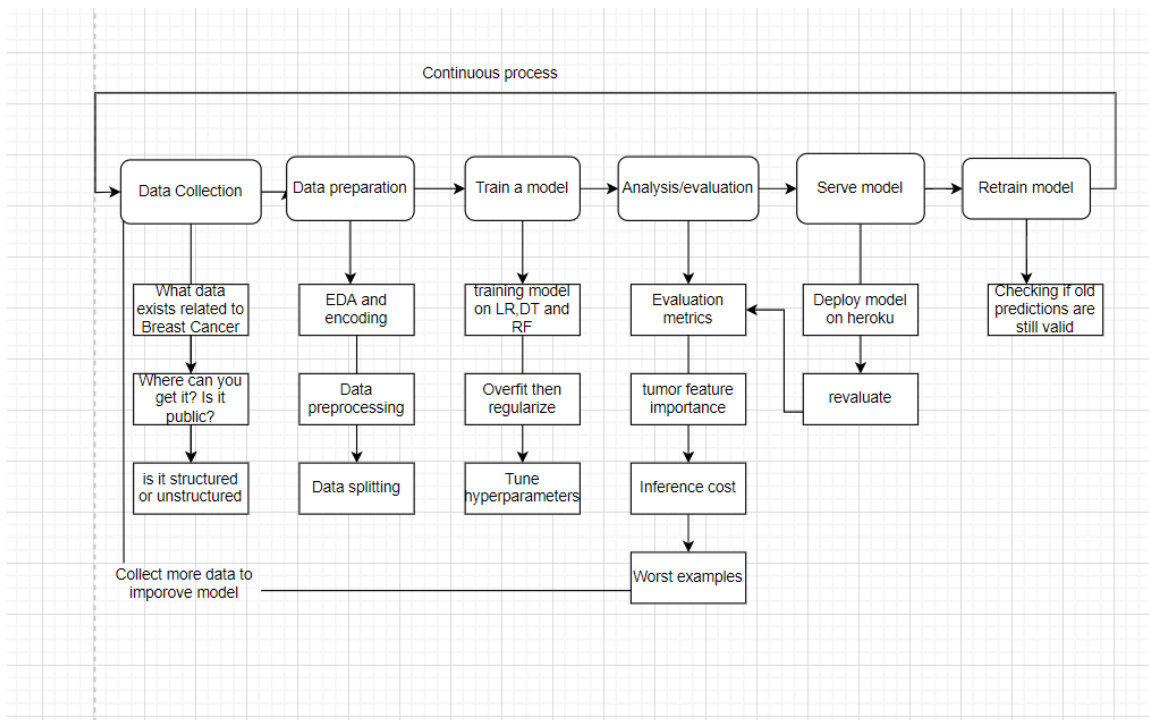


Figure 3.1: High Level Design

## 3.2 Detailed Design

Streamlit is going backwards by mixing the code and the layout in a single entity. While, that is true, it is also attempting to make life easier because all of those technologies mentioned above are no longer necessary — all you need is Python and the Streamlit library.

The solution is to keep the MVC approach in mind and separate the program logic from the way it is presented.



Figure 3.2: Detailed Design

### 3.3 Sequence Design

As shown in the sequence diagram for training flow, Pathologist invokes train and trains a Random Forest model based on the training dataset. As shown in the sequence diagram for Prediction, Random Forest invokes predict whether it is Benign or Malignant[3].

The sequence diagram corresponding to the project is depicted in Figure below:

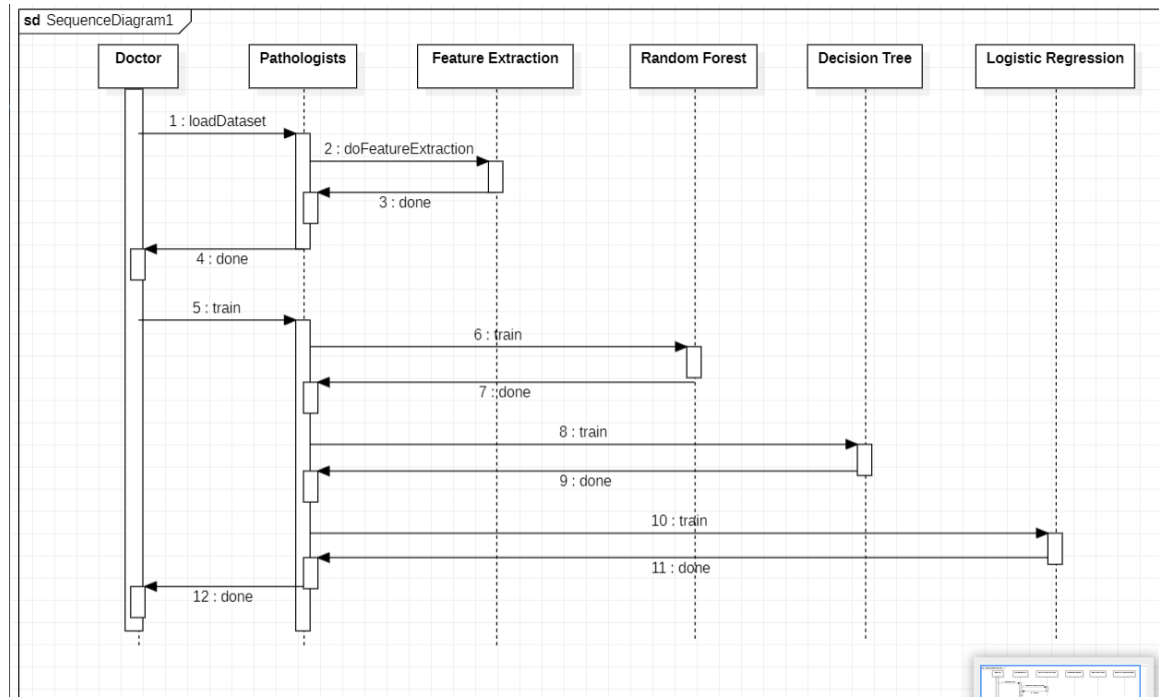


Figure 3.3: Sequence Design

### 3.4 Usecase Design

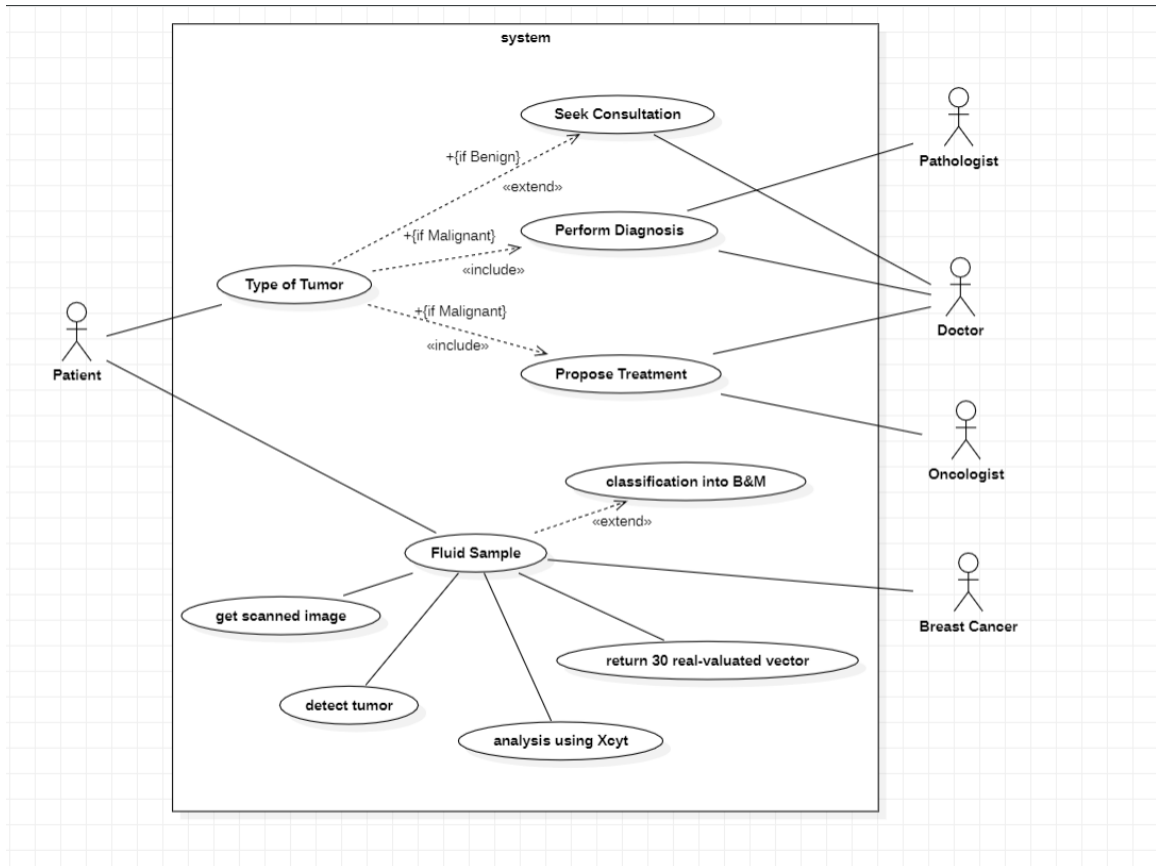


Figure 3.4: Usecase Design

#### Usecase:

A use case diagram is usually simple. It does not show the detail of the use cases:

- It only summarizes some of the relationships between use cases, actors, and systems.
- It does not show the order in which steps are performed to achieve the goals of each use case[4].



The use-case diagram corresponding to the project is depicted in the figure below  
There are five users:

1. Patient
2. Pathologist
3. Doctor
4. Oncologist
5. Breast cancer system

## Chapter 4

# Implementation

### 4.1 Proposed methodology

The goal is to classify whether the breast cancer is benign or malignant. To achieve this we have used machine learning classification methods to fit a function that can predict the discrete class of new input. The methodology followed includes the following phases:

Phase 1 — Data Exploration

Phase 2 — Categorical Data

Phase 3 — Feature Scaling

Phase 4 — Model Selection

We will use Classification Accuracy method to find the accuracy of our models. Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples[5].

### 4.2 Algorithm used for implementation

#### 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 predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

M12.model 2				
	precision	recall	f1-score	support
...				
macro avg	0.96	0.97	0.96	57
weighted avg	0.97	0.96	0.97	57
Accuracy: 0.9649122807017544				

Figure 4.1: Logistic Regression Classification Report

### Decision tree

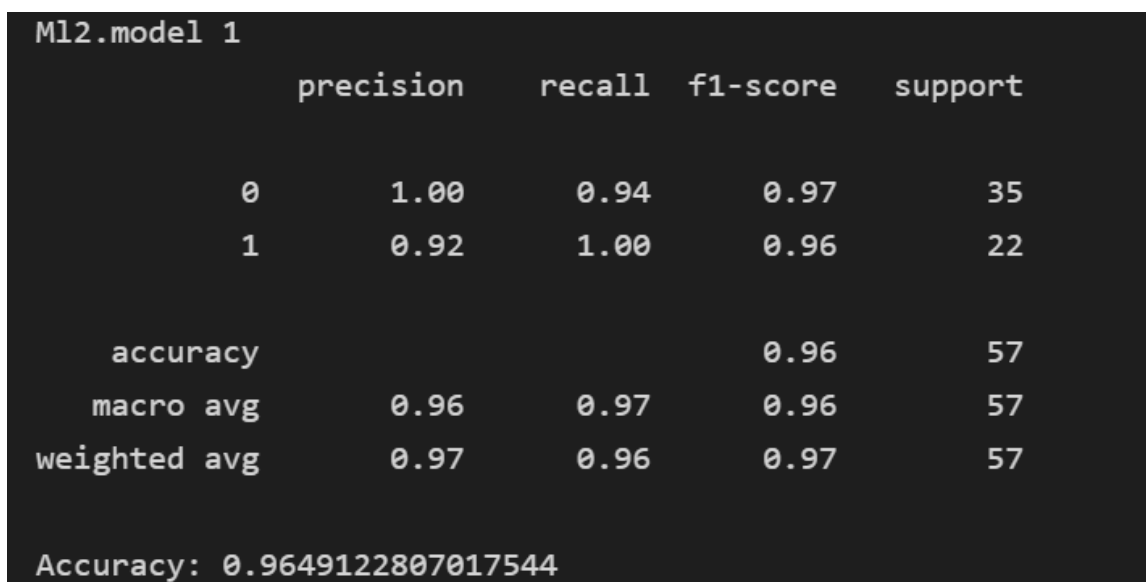
Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. 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[6].

M12.model 0				
	precision	recall	f1-score	support
0	0.97	0.91	0.94	35
1	0.88	0.95	0.91	22
accuracy			0.93	57
macro avg	0.92	0.93	0.93	57
weighted avg	0.93	0.93	0.93	57

Figure 4.2: Decision Tree Classification Report

## 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. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "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[7].



```
Ml2.model 1
              precision    recall  f1-score   support

      0         1.00        0.94        0.97         35
      1         0.92        1.00        0.96         22

   accuracy                0.96         57
  macro avg         0.96        0.97        0.96         57
weighted avg         0.97        0.96        0.97         57

Accuracy: 0.9649122807017544
```

Figure 4.3: Random Forest Classification Report

## 4.3 Tools and technologies used

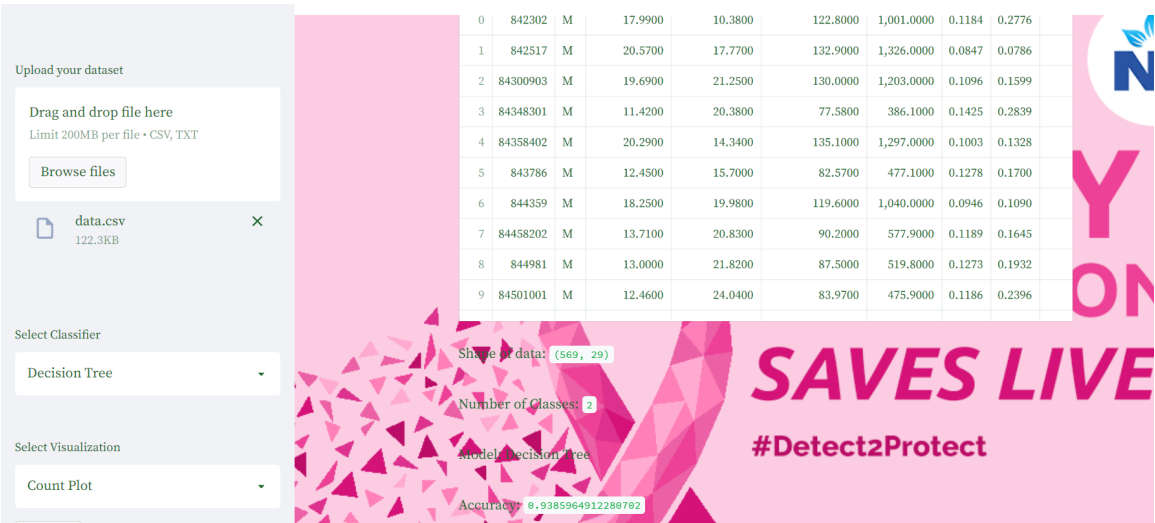
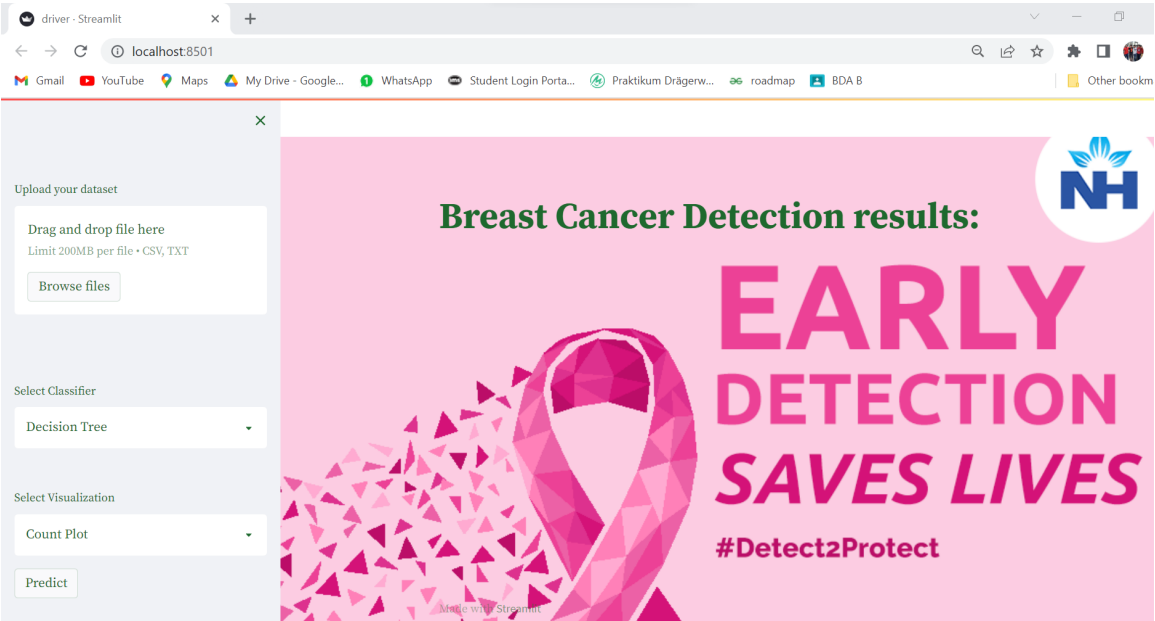
**Front end app framework:** Stream lit

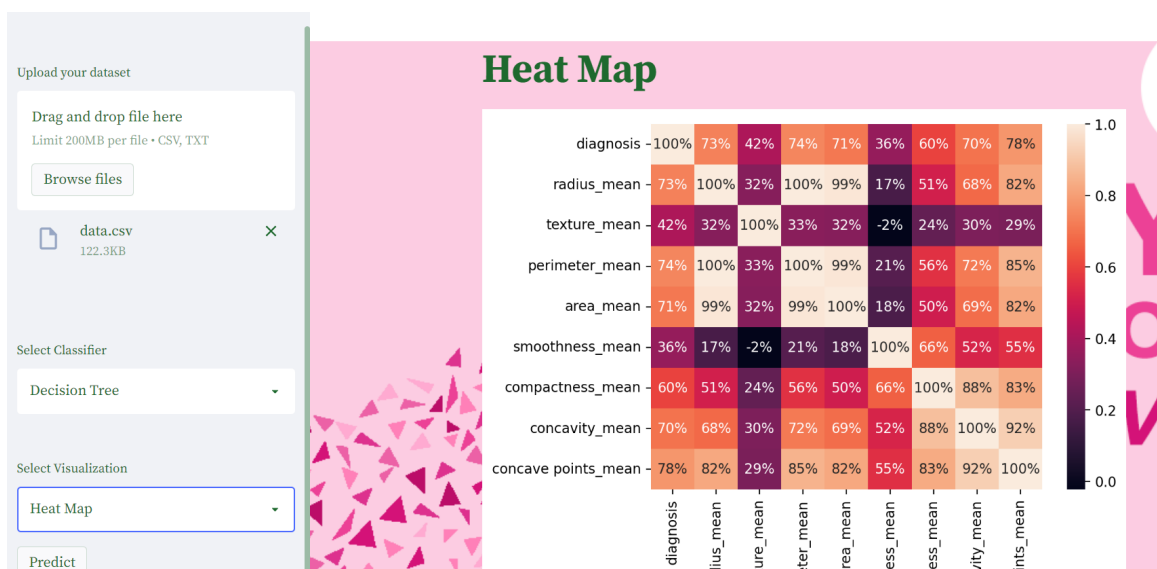
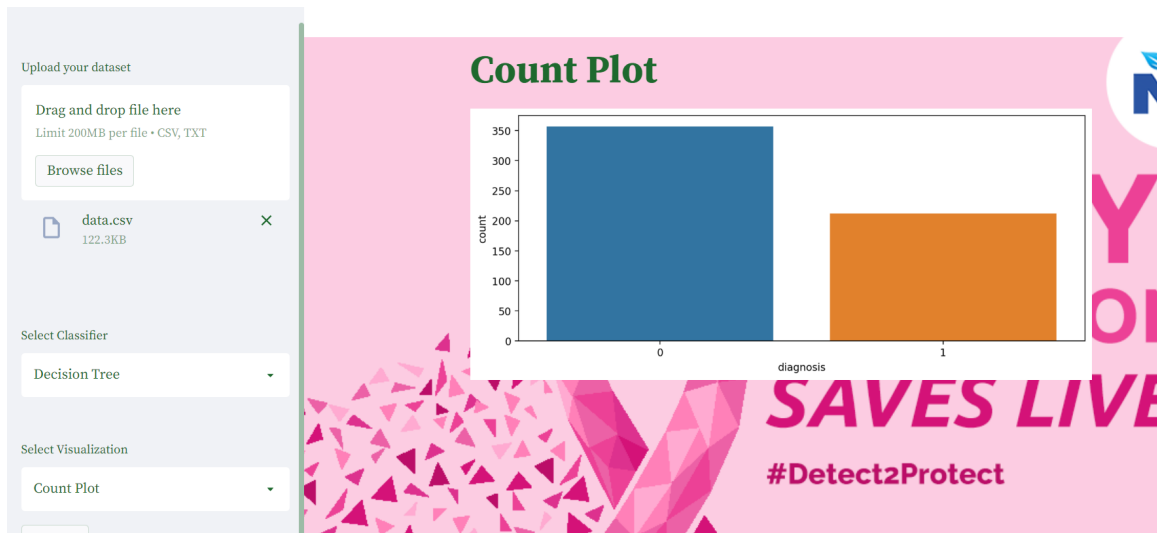
**Web framework:** Flask

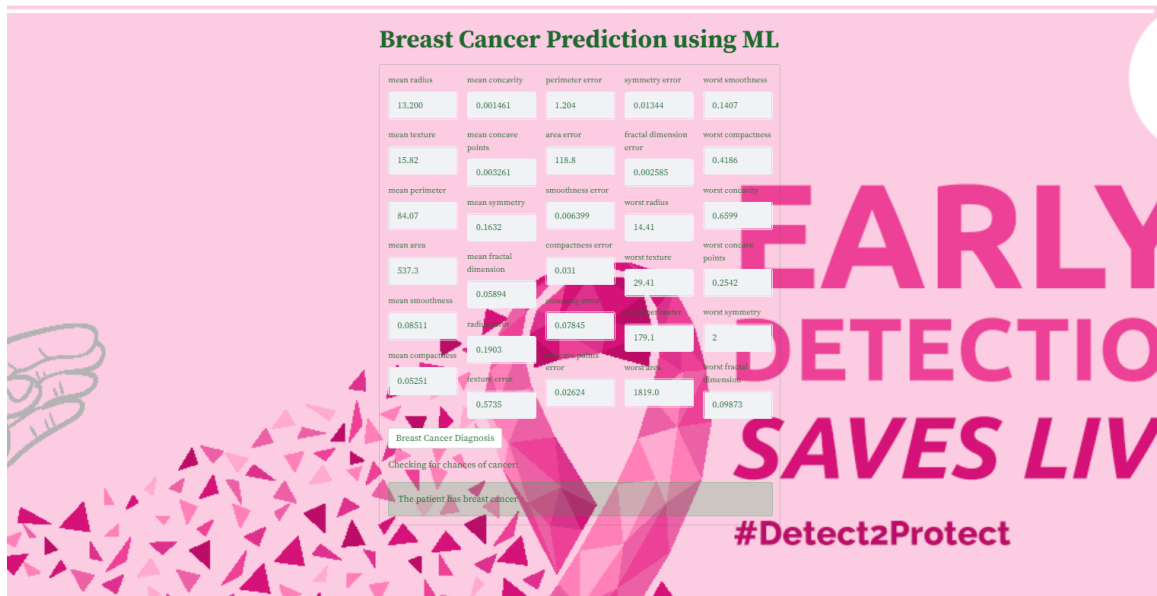
**Language:** Python

**Libraries:** scikit-learn, pandas, numpy, seaborn, matplotlib

# 4.4 Screenshots







## 4.5 Testing

In this project ,we test our datasets with different algorithms and we predict which algorithm works better and faster for our dataset .we calculate the accuracy of the algorithms and compare them with other algorithms . The best algorithm which works on this dataset is taken and we are predicting the factors that are causing the employee attrition.

## Chapter 5

# Results and Discussion

Out of all algorithms it was found that decision tree gave highest accuracy of 1.0. However this is a case of overfitting. Over-fitting occurs when the tree is designed so as to perfectly fit all samples in the training data set. Thus it ends up with branches with strict rules of sparse data. Thus this effects the accuracy when predicting samples that are not part of the training set[8]. One of the methods used to address over-fitting in decision tree is called pruning which is done after the initial training is complete. In pruning, you trim off the branches of the tree, i.e., remove the decision nodes starting from the leaf node such that the overall accuracy is not disturbed. This is done by segregating the actual training set into two sets: training data set, D and validation data set, V. Prepare the decision tree using the segregated training data set D. Then continue trimming the tree accordingly to optimize the accuracy of the validation data set V. Random Forest was found to have an accuracy of 0.998046875. Random forests overcome several problems with decision trees, including:

- Reduction in overfitting: by averaging several trees, there is a significantly lower risk of overfitting.
- Less variance: By using multiple trees, you reduce the chance of stumbling across a classifier that doesn't perform well because of the relationship between the train and test data.
- It can also maintain accuracy when a large proportion of data is missing[9].

As a result in almost all cases, random forests are more accurate than decision trees. Hence we shall be concluding by stating random forest to be the best performing algorithm.



```
model=models(X_train,Y_train) #tuple of all models
✓ 0.4s

1. Decision Tree accuracy: 1.0
2. Random Forest accuracy: 1.0
3. Logistic Regression accuracy: 0.9621212121212122
```

Figure 5.1: Result After Dropping 120 Rows

```
✓ 0.7s

1. Decision Tree accuracy: 1.0
2. Random Forest accuracy: 0.9978021978021978
3. Logistic Regression accuracy: 0.9912087912087912

Test results
```

Figure 5.2: Result After Dropping 2 Columns

```
model=models(X_train,Y_train) #tuple of all models
✓ 0.6s

1. Decision Tree accuracy: 1.0
2. Random Forest accuracy: 0.998046875
3. Logistic Regression accuracy: 0.986328125
```

Figure 5.3: Result After Dropping 15 Columns

## 5.1 ROC

A Receiver Operator Characteristic (ROC) curve is a graphical plot used to show the diagnostic ability of binary classifiers. It was first used in signal detection theory but is now used in many other areas such as medicine, radiology, natural hazards and machine learning. A ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR). The true positive rate is the proportion of observations that were correctly predicted to be positive out of all positive observations ( $TP/(TP + FN)$ ). Similarly, the false positive rate is the proportion of observations that are incorrectly predicted to be positive out of all negative observations ( $FP/(TN + FP)$ ). For example, in medical testing, the true positive rate is the rate in which people are correctly identified to test positive for the disease in question[10].

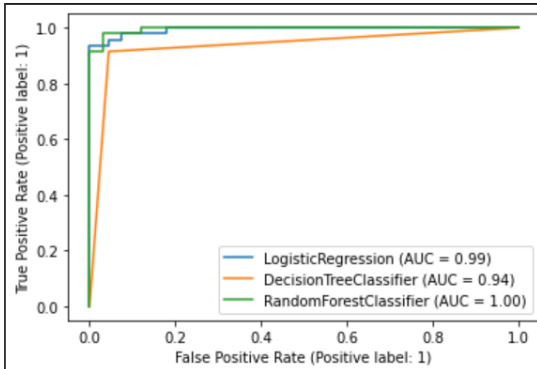


Figure 5.4: ROC Curve

## Chapter 6

# Conclusion and Future Work

In this paper we examined different machine learning techniques for breast cancer detection. We performed a comparative analysis of Decision Tree, Random Forest and Logistic regression. It was observed that Random Forest outperforms the existing methods on issues related to accuracy, precision and also size of the data set. As future work on this project we plan to apply boosting. Boosting is an ensemble modelling, technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added. XGBoost is an implementation of Gradient Boosted decision trees. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model.

# Bibliography

- [1] S. Sharma, A. Aggarwal, and T. Choudhury, “Breast cancer detection using machine learning algorithms,” *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pp. 114–118, 2018.
- [2] L. Hussain, W. Aziz, S. Saeed, S. Rathore, and K. Rafique Mir, “Automated breast cancer detection using machine learning techniques by extracting different feature extracting strategies,” 08 2018, pp. 327–331.
- [3] A. Meriem, S. Oukid, I. Gagaoua, and T. Ensari, “Breast cancer classification using machine learning,” 04 2018, pp. 1–4.
- [4] M. Tahmooresi, A. Afshar, B. B. Rad, K. B. Nowshath, and M. A. Bamiah, “Early detection of breast cancer using machine learning techniques,” *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 10, pp. 21–27, 2018.
- [5] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, “Machine learning classification techniques for breast cancer diagnosis,” in *IOP Conference Series: Materials Science and Engineering*, vol. 495, no. 1. IOP Publishing, 2019, p. 012033.
- [6] M. Islam, M. Haque, H. Iqbal, M. Hasan, M. Hasan, M. N. Kabir *et al.*, “Breast cancer prediction: a comparative study using machine learning techniques,” *SN Computer Science*, vol. 1, no. 5, pp. 1–14, 2020.
- [7] H. Aljuaid, N. Alturki, N. Alsubaie, L. Cavallaro, and A. Liotta, “Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning,” *Computer Methods and Programs in Biomedicine*, p. 106951, 2022.
- [8] H. Saoud, A. Ghadi, and M. Ghailani, “Proposed approach for breast cancer diagnosis using machine learning,” in *Proceedings of the 4th international conference on smart city applications*, 2019, pp. 1–5.

- [9] S. S. M. Khairi, M. A. A. Bakar, M. A. Alias, S. A. Bakar, C.-Y. Liong, N. Rosli, and M. Farid, “Deep learning on histopathology images for breast cancer classification: a bibliometric analysis,” in *Healthcare*, vol. 10, no. 1. MDPI, 2021, p. 10.
- [10] A. Kajala and V. Jain, “Diagnosis of breast cancer using machine learning algorithms-a review,” in *2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3)*. IEEE, 2020, pp. 1–5.