# PREDICTION OF BREAST CANCER USING MACHINE LEARNING

### A PROJECT REPORT

#### Submitted by

HARSH PANDEY (22BCS17030)

ARYADEEP (22BCS10915)

ROSHAN LAL NATHANI (22BCS16775)

HARMAN SINGH (22BCS14925)

MOHAN KARKI (22BCS17035)

***in partial fulfillment for the award of the degree of***

**BACHELOR OF ENGINEERING**

**IN**

COMPUTER SCIENCE ENGINEERING



**Chandigarh University**

MARCH 2023

**TABLE OF CONTENTS**

[List of Figures](#_TOC_250001)

[List of Tables](#_TOC_250000)

List of Standards

**CHAPTER 1. INTRODUCTION 3**

* 1. [Identification of Client/ Need/ Relevant Contemporary issue 3](#_bookmark0)
  2. [Identification of Problem 3](#_bookmark1)
  3. [Identification of Tasks 4](#_bookmark2)
  4. [Timeline 5](#_bookmark3)
  5. [Organization of the Report 5](#_bookmark4)

**REFERENCES**

**INTRODUCTION**

* 1. **Identification of Client /Need / Relevant Contemporary issue/Project Scope**

Breast Cancer is the most widely recognized type of cancer and is the main kind of malignant growth that is boundless among ladies around the world. As indicated by the report of WHO (World Health Organization) every year, 2.1 million ladies are recently affected by breast cancer.

Furthermore, there should be a proper prediction and classification about the type of cancer a patient is having so doctor may know whether surgery is required or not. Such classification will help doctor to do surgeries only when the cancer is malignant, thus reducing the unnecessary surgeries.

The dataset used in this model was created by Dr. William H. Wolberg, Physician at the University of Wisconsin, USA.

The project is about building a Logistic Regression Model using Sklearn to classify breast cancer as either Malignant or Benign.

Logistic Regressionis one of the best models when it comes to binary classification as we are going to classify the data points into two types here.

The two types of classified data which we are using in this model, and they are benign (1) and malignant (0).

So, once you train your Logistic Regression model with this particular Training dataset, we will have a trained Logistic Regression model and we will also do some evaluation of this model using Teat data. After that we can have a new data and once you give this new data the model can tell you whether that particular tumour is benign or malignant.

* 1. **Identification of Problem**

Breast cancer is the most common type of cancer in women. When cancers are found early, they can often be cured. There are some devices that detects the breast cancer but many times they lead to false positives, which results in patients undergoing painful, expensive surgeries that were not even necessary. These types of cancer are called **Benign** which do not require surgeries and we can reduce these unnecessary surgeries by using Machine Learning. We take a classified dataset of the previous breast cancer patients and train the model to predict whether the cancer is **benign**or**malignant**. These classifications will help doctors to do surgeries only when the cancer is malignant, thus reducing the unnecessary surgeries for woman.

* 1. **Identification of Tasks**

As you know, machines initially learn from the data that you give them. It is of the utmost importance to collect reliable data so that your machine learning model can find the correct patterns. The quality of the data that you feed to the machine will determine how accurate your model is. If you have incorrect or outdated data, you will have wrong outcomes or predictions which are not relevant.

After you have your data, you have to prepare it. You can do this.

Putting together all the data you have and randomizing it. This helps make sure that data is evenly distributed, and the ordering does not affect the learning process.

Cleaning the data to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc. You might even have to restructure the dataset and change the rows and columns or index of rows and columns.

Visualize the data to understand how it is structured and understand the relationship between various variables and classes present.

Splitting the cleaned data into two sets - a training set and a testing set. The training set is the set your model learns from. A testing set is used to check the accuracy of your model after training.

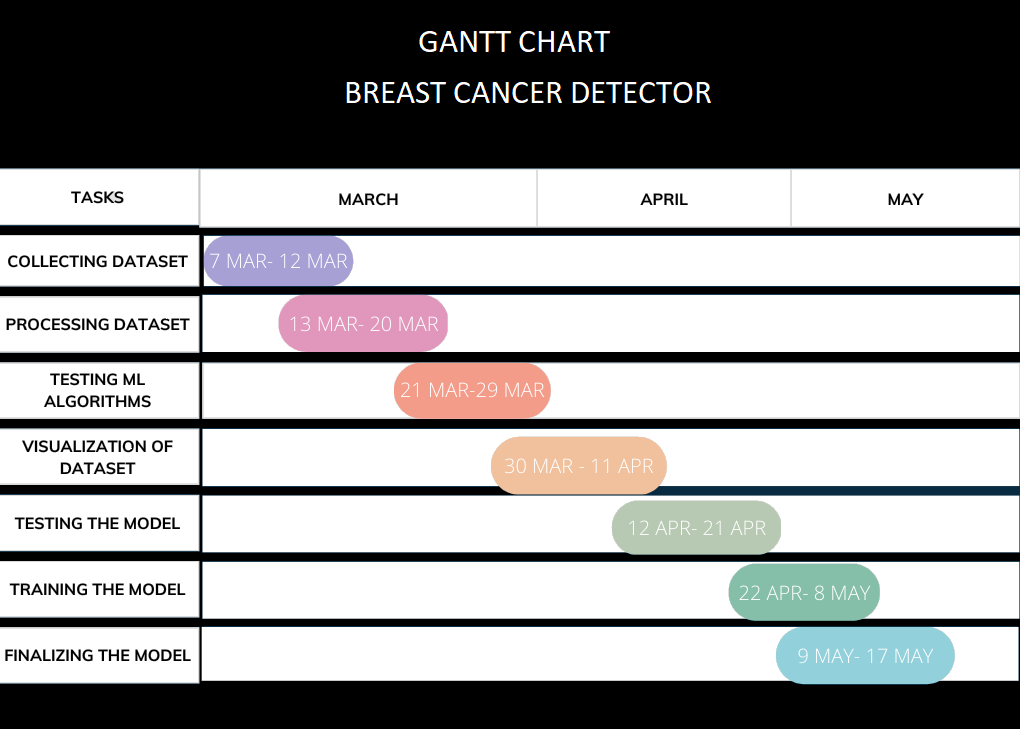
A machine learning model determines the output you get after running a machine learning algorithm on the collected data. It is important to choose a model which is relevant to the task at hand. Over the years, scientists and engineers developed various models suited for different tasks like speech recognition, image recognition, prediction, etc.

Training is the most important step in machine learning. In training, you pass the prepared data to your machine learning model to find patterns and make predictions. It results in the model learning from the data so that it can accomplish the task set. Over time, with training, the model gets better at predicting.

High accuracy After training your model, you have to check to see how it’s performing. This is done by testing the performance of the model on previously unseen data. The unseen data used is the testing set that you split our data into earlier. If testing was done on the same data which is used for training, you will not get an accurate measure, as the model is already used to the data, and finds the same patterns in it, as it previously did. This will give you disproportionately.

Once you have created and evaluated your model, see if its accuracy can be improved in any way. This is done by tuning the parameters present in your model. Parameters are the variables in the model that the programmer generally decides. At a particular value of your parameter, the accuracy will be the maximum. Parameter tuning refers to finding these values.

* 1. **Timeline**

****

* 1. **Organization of the Report**

**Chapter 1 Problem Identification:** This chapter introduces the project and describes the problem statement discussed earlier in the report.

**Chapter 2 Literature Review:** This chapter prevents review for various research papers which help us to understand the problem in a better way. It also defines what has been done to already solve the problem and what can be further done.

**Chapter 3 Design Flow/ Process:** This chapter presents the need and significance of the proposed work based on literature review. Proposed objectives and methodology are explained. This presents the relevance of the problem. It also represents logical and schematic plan to resolve the research problem.

**Chapter 4 Result Analysis and Validation:** This chapter explains various performance parameters used in implementation. Experimental results are shown in this chapter. It explains the meaning of the results and why they matter.

**Chapter 5 Conclusion and future scope:** This chapter concludes the results and explain the best method to perform this research to get the best results and define the future scope of study that explains the extent to which the research area will be explored in the work.