A Mini Project Report

*On*

**Breast Cancer Prediction**

*by*

Tejaswini Ugale (T194108)

Tanmay Kulkarni (T194127)

Shoyab Shaikh (T194128)

Shantanu Kukade (T194129)

*of*

**T.Y. B.Tech**

**Academic Year: 2019-2020**

**School of Computer Engineering & Technology**



**Contents**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No.** | **Topic** | | **Page No.** |
| **1** | **Aim & Objective** | | iii |
| **2** | **Chapter-1 Introduction** | |
|  | 1.1 | Introduction to Project |
| 1.2 | Problem Statement |
| **3** | **Chapter-2 Survey** | | iv |
|  | 2.1 | Background Study of the project topic |
| **4** | **Chapter-3 Proposed Model** | | v |
|  | 3.1 | Algorithm used |
| **5** | **Chapter-4 Dataset** | | vi- vii |
|  | 4.1 | Details of dataset |
| 4.2 | Feature vector used |
| 4.3 | Graph representation of data set |
| **6** | **Chapter-5 Results & Analysis (Graph visualization)** | | viii |
| **7** | **Chapter-6 Conclusion & Future work** | | ix |
|  | **References** | | ix |

**Aim**

* The goal is to classify whether the breast cancer is benign or malignant using machine learning.

**Chapter 1 Introduction**

**1.1 Introduction to the Project**

* 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. It get developed from breast tissue, signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, a newly-inverted nipple, or a red or scaly patch of skin. Its treatment depends on the stage of cancer. It may consist of chemotherapy, radiation, hormone therapy and surgery.
* The early diagnosis of Breast cancer can improve the prognosis and chance of survival significantly, as it can improve timely clinical treatment of patients. Further accurate classification of Benign and malignant tumors can prevent undergoing unnecessary treatments where machine learning is most widely recognized methodology for classification of Breast cancer.

**1.2 Problem Statement**

To build an intelligent system that will predict breast cancer for a patient and will help doctors using suitable machine learning algorithm.

**Chapter 2 Survey**

* **2.1 Background Study of the project topic**
* **Mammography**
* The most important screening test for breast cancer is the mammogram. It can detect breast cancer up to two years before the tumor can be felt by you or your doctor. A mammogram is an x-ray picture of the breast. It can be used to check for breast cancer in women who have no signs or symptoms of the disease. It can also be used if you have a lump or other sign of breast cancer. Screening mammography is the type of mammogram that checks you when you have no symptoms.

**Breast Cancer**

* 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. It get developed from breast tissue, signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, a newly-inverted nipple, or a red or scaly patch of skin. Its treatment depends on the stage of cancer. It may consist of chemotherapy, radiation, hormone therapy and surgery.
* The early diagnosis of Breast cancer can improve the prognosis and chance of survival significantly, as it can improve timely clinical treatment of patients. Further accurate classification of Benign and malignant tumors can prevent undergoing unnecessary treatments where machine learning is most widely recognized methodology for classification of Breast cancer.

**Chapter 3 Proposed Model.**

**3.1 Algorithm used**

The algorithm used for training the data is Logistic Regression comes into classification category of supervised learning, machine learning. Logistic Regression is a supervised machine learning technique, employed in classification jobs. Binary outcomes can be predicted from the independent variables. The outcome of dependent variable is discrete. Logistic Regression uses a simple equation which shows the linear relation between the independent variables.

This algorithm is entitled as logistic regression as the key method behind it is logistic function. The output can be predicted from the independent variables, which form a linear equation. The output predicted has no restrictions; it can be any value from negative infinity to positive infinity. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

**Chapter 4 Dataset**

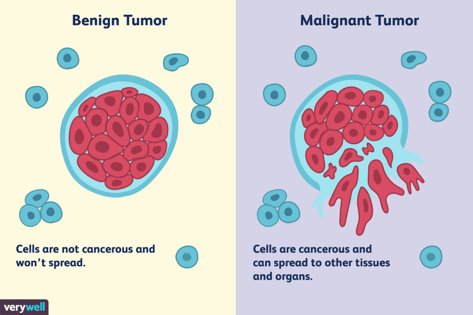
**4.1 Details of the dataset**

* The dataset used for our system is the UCI Machine Learning Repository for breast cancer dataset. It contains various columns which need to be considered for breast cancer. The data is huge and contains around 500 records with 31 columns. The dataset contains all records which define the properties of a breast , it contains a unique id associated to each record.
* **4.2 Feature Vector**
* The dataset contains following field which are taken into consideration for training the dataset.

1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter² / area — 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension (“coastline approximation” — 1)

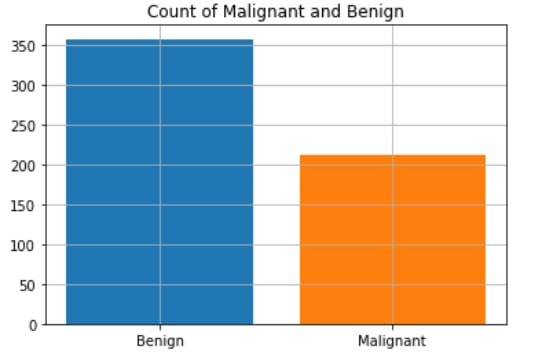
Dependent Variable

The dependent variable here is the diagnosis field which tells us that whether a given cancer is *benign* or *malignant*, breast cancer is classified into two categories : benign cancer and malignant cancer which is given as follows



**4.3 Graph Representation of dataset**

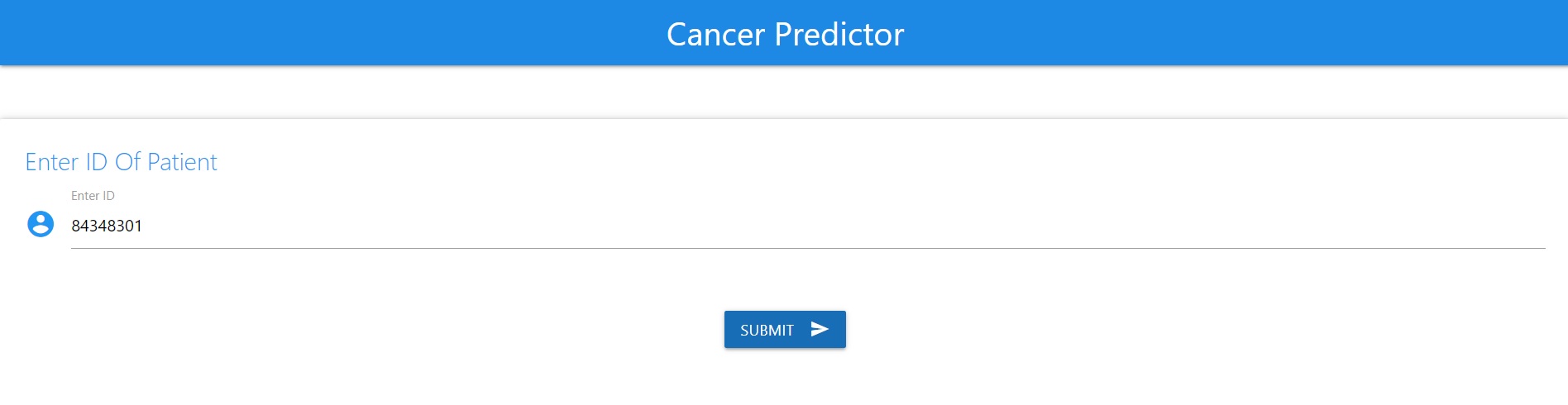
The following graph is the count of total number of patients having a type benign breast cancer and total number of patient having malignant type cancer in the given UCI datatset.



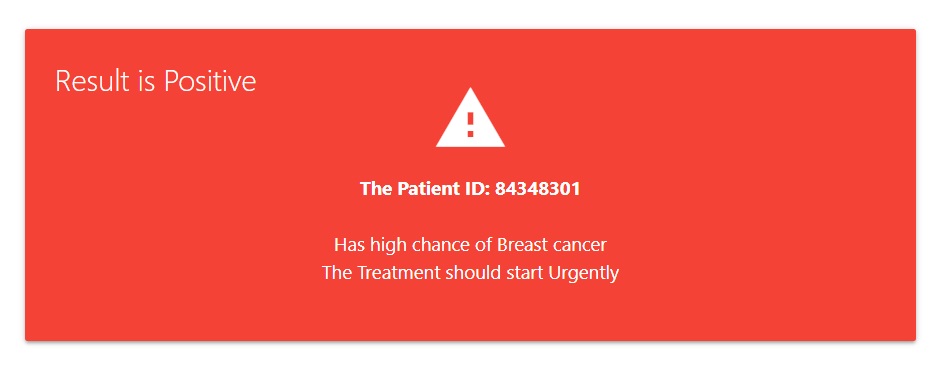
**Chapter 4 Results & Analysis**

According to the results and analysis the system predicts the cancer for a given patient by considering his/her mammogram and thus classifies whether the cancer is benign or malignant.

The doctor will just need enter the patient id and he can view the various mammogram readings for the given patient and he then check for the chances of cancer.



****

****

**Chapter 6 Conclusion and Future Work:-**

Results conclude that Simple Logistic regression method obtains the Best Model to predict breast cancer by means of different data mining techniques. Results indicate that Simple Logistic regression obtained best performance in general compared to the other classifiers in terms of classification accuracy. Also the system truly provide desired redults by predicting the breast cancer and performs efficiently.

**References**

1. <https://www.verywellhealth.com/>
2. [https://archive.ics.uci.edu](https://archive.ics.uci.edu/)
3. UCI Machine Learning Repository,: <http://archive.ics.uci.edu/ml/>
4. McCarthy et al. Applications of Machine Learning and High -Dimensional Visualization in Cancer Detection, Diagnosis, and Management.