Term Paper

On

BREAST CANCER SURVIVAL PREDICTION SYSTEM USING MACHINE LEARNING

Submitted to

Amity School Of Engineering And Technology

Guided by

Dr. MONIKA ARORA

Submitted by

NIKITA

A023119820034

DEPARTMENT OF ARTIFICIAL INTELLIGENCE

AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY

GAUTAM BUDDHA NAGAR NOIDA UTTAR PRADESH

STUDENT DECLARATION

I, **Nikita**, student of Bachelor's in technology in Artificial Intelligence department hereby declare that the project titled **Breast Cancer Survival Prediction System Using Machine Learning**, which is submitted by me to the Department of Artificial Intelligence at Amity School of Engineering and Technology. Noida, Uttar Pradesh. In partial fulfillment of requirement for award of the degree of Bachelor of Technology in Artificial Intelligence has not been previously formed the basis for the reward of any degree, diploma, or other similar title work recognition.

The author attests that permission has been obtained from the use of any copyrighted material appearing in the report other than brief experts requiring only proper acknowledgement in the scholarly writing and all such use is acknowledged.

Marwat

NOIDA

DATE: 20 JUNE 2022

CERTIFICATE

On the basis of the report submitted by **Nikita**, student of Bachelor's in technology in Artificial Intelligence, I hereby certify that the report on **Breast Cancer Survival Prediction System Using Machine Learning**, which is submitted to the Department of Artificial Intelligence ,Amity School of Engineering and Technology, Uttar Pradesh In partial fulfillment of the requirements of for the award of the Degree of Bachelor of Technology is an original contribution with the existing knowledge and faithful record of work carried out by her Under my guidance and supervision to the best of my knowledge this work has not been submitted in part or full for any degree or diploma to this university or anywhere else.

NIKITA

20 JUNE 2022

Dr. MONIKA ARORA

Faculty Guide

Department of Engineering and Technology

ACKNOWLEDGEMENT

I would like to express my deep gratitude towards almighty god without his blessings I was not be able to complete this paper.

This paper and the research behind it would not have been possible without the exceptional support of my faculty guide, Dr, Monika Arora, Department of Engineering and technology, Amity School of Engineering and Technology, Noida, Uttar Pradesh.

Her enthusiasm, knowledge and meticulous attention to detail inspired me to do my job well.

I also want to thank my program leader, Dr. Sanjay Dubey, Department of Engineering and technology, Amity School of Engineering and Technology, Noida, Uttar Pradesh. who helped in the initial stages of my paperwork and also guided me with the do’s and don’ts of the registration process.

I also appreciate the insightful comments from our anonymous colleagues. Everyone's generosity and expertise have strengthened this research in myriad ways and saved me from many mistakes, inevitably remain is entirely my own responsibility.

Marwat

INDEX

ABSTRACT

Breast cancer is the third most dangerous type of cancer taking large number of lives every day. So, it becomes mandatory for a system to give most reliable and accurate results in this sphere. Breast Cancer Survival Prediction System will help doctors to reach to correct decision more efficiently.

Python language is used for machine learning for this framework. Breast cancer prediction system will entirely work on Logistic Regression framework to foresee the survival possibilities of the breast cancer patient after surgery.

This model or framework will envision the after result of the surgery by studying patterns and relationship in the dataset stored of carcinoma patients.

Breast Cancer Prediction system will help the physicians to predict that after which type of surgery i.e., (Lumpectomy, Simple Mastectomy, Modified Radical Mastectomy or Other) the patient will be dead or alive.

It ought to be brought into thought that unlike patients have different parameters (Protein1, Protein2, Protein3, Protein4, Histology, ER status, PR status, HER2 status) and different stages (Tumour Stage) of breast cancer.

In the end of the project, one is able to understand the real life working of regression model and able to learn how to train a model and test it as well.

INTRODUCTION

CHAPTER 1:

Cancer is probably most feared and frightening illnesses among all diseases. In accordance with World Health Organization, it is the second leading cause of death. This type of carcinoma not only affects feminine but masculine too. However, the rate or proportion of breast cancer is more among females as compared to males. Machine learning is applied to foresee whether breast cancer patient will survive after surgery or not. Machine learning is widely using datasets to help in healthcare.

1.1 This task is performed with the help machine learning in python language. In this project a dataset sample of over 350 carcinoma affected role have underwent surgery is used which is collected from Kaggle. Logistic regression model under supervised learning is used for prediction of chances of survival for cancer patient post-surgery.

- **1.1.1** Supervised learning is machine learning, where a machine is trained with T.D and the machine predicts output based on that data. Tagged data means that some input data will be used with the correct output.

  In supervised learning, the training data provided to system acts as a supervisor, instructing the machine to predict the output correctly.

- **1.1.2** A Logistic Regression model is a statistical model that demonstrates the possibility of an incident (of two choices) will occur through taking log odds of the event as a linear combination of one or more independent variables.

1.2 Numerous alternative models may also be used in place of current one for this prediction model like decision tree, random forest, neural networks, extreme boost, and support vector machine but only logistic regression model is discussed in this paper as it is easier to understand and gives reliable results.

Before beginning with our prediction model lets discuss about the dataset, we are going to use in this prediction model.

DATASET DISCUSSION

CHAPTER 2:

The dataset used in this prediction model is taken from Kaggle. In this dataset data of over 350 cancer patients is used who underwent surgery. This dataset forms __ number of images. Now let’s discuss all the fields given in the data set:

2.1

FIELD_NAME	DISCRIPTION
Patient_ID	Unique ID of every patient through which s/he is identified.
Age	Patient’s age
Gender	Patient’s gender (M/F)
Protein1, Protein2, Protein3, Protein4	Undefined expression levels
Tumour_Stage	Patient’s stage of cancer
Histology	Microscopic structure of tissues (Infiltrating Ductal Carcinoma, Infiltration Lobular Carcinoma, Mucinous Carcinoma)
ER status	Hormone receptors status (Negative/ Positive)
PR status	Progesterone receptors status (Negative/Positive)
HER2 status	Negative/Positive
Surgery_type	Type of surgery patient underwent (Lumpectomy, Simple Mastectomy, Modified Radical Mastectomy or Other)
Date_of_Surgery	Appointment of surgery
Date_of_Last_Visit	Patient last visited date
Patient_Status	Whether the patient is dead or alive

Table 2.1 showing description of the dataset used

PREDICTION MODEL PYTHON CODE DISCUSSION

CHAPTER 3:

➢ 3.1 LIBRARIES USED:

1) PANDAS:

   With Pandas, you can analyze big data & illustrate deductions centered on statistical theory. Pandas can clean up cluttered datasets to make them easier to read and relevant.

2) NUMPY:

   NumPy is a general-purpose array handling package. It provides high-running M-dimensional array objects and tools for manipulating those arrays. This is the basic package of Python scientific calculations.

3) PLOTLY:

   Plotly allows users to interact with the graphics displayed, improving the storytelling experience. It is basically employed for generating charts and displaying statistics in a interactive manner.

4) SCIKIT LEARN:

   Sklearn is extensively used for classification, predictive analytics, and numerous other machine learning tasks. It is very approachable and powerful library.

5) TRAIN TEST SPLIT:

   Train test split techniques are used to estimate the performance of ML algos which are used at the time of prophecy of statistics which are not used to train a model. This technique is suitable if your dataset is exceptionally large, if you have an expensive model for training, or if you need to quickly estimate the performance of your model.

6) SUPPORT VECTOR CLASSIFIER (SVC):

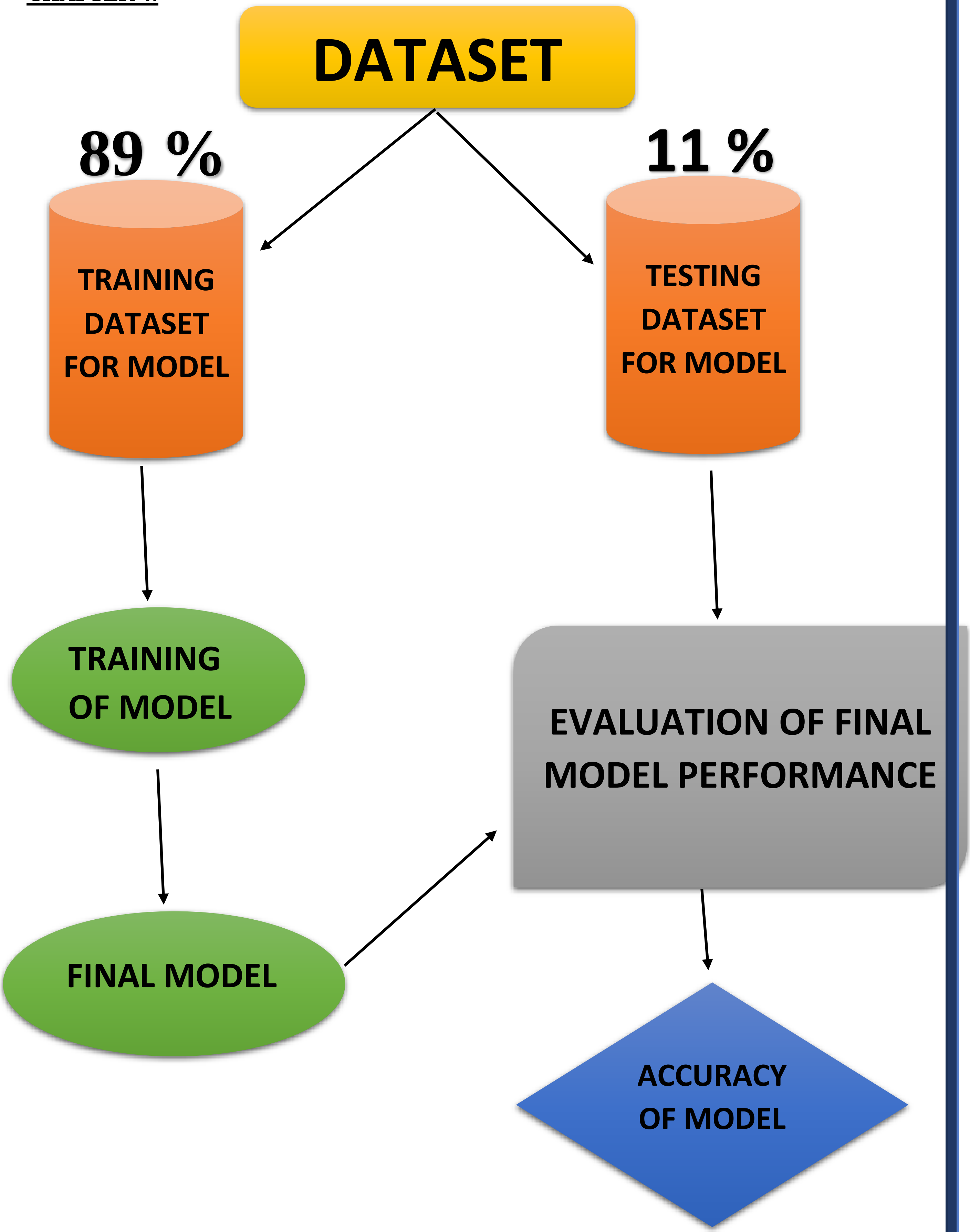
   The goal of a SVC is to fit the numbers you deliver and return an "optimal" hyperplane that subdivides or classifies data. All of these are making this algo best suited for use, but this could be employed in many situations.

➢ 3.2 TOOL USED:

Google Colab Notebook

BLOCK DIAGRAM

CHAPTER 4:


DATASET
89 %
11 %
TRAINING DATASET FOR MODEL
TESTING DATASET FOR MODEL
TRAINING OF MODEL
EVALUATION OF FINAL MODEL PERFORMANCE
FINAL MODEL
ACCURACY OF MODEL

PROPOSED METHODOLOGY

CHAPTER 5:

5.1 The unit diagram discussed above clearly shows the working of the breast cancer survival prediction model. In this section we will discuss the methodology of the model briefly.

All the libraries discussed previously are imported for the model.

Dataset is then checked whether it contains any NULL values or not. Each of the units comprising the NULL values within the dataset are dropped.

Proportion of males and females having breast cancer out of total dataset is calculated.

A circular chart is created depicting the stages of tumor of the breast cancer patients. Stages are named as I, II, and III.

Histology of the patients is brought into concern w.r.t how quickly malignant cells can develop and spread among Infiltrating Ductal Carcinoma, Infiltration Lobular Carcinoma, and Mucinous Carcinoma. A circular chart is created to show the distribution.

ER (Estrogen receptor) status, PR (Progesterone receptors) status, and HER2 status of the patients are considered.

Another circular chart is created to show the percentage of patients have undergone which type of surgery (Lumpectomy, Simple Mastectomy, Modified Radical Mastectomy or Other).

After exploration of the data, it is found to have categorical features. In order to work out the ML framework by this data, all categorical columns values are transformed.

To start the training of the ML framework in order to find the prospect of the patients' survival, the dataset is divided into two subsets i.e., training set & test set.

5.2 Eventually all the features are keyed in to find the survival probability of the cancer patient.

- TRAINING SET: It is the subset for training the model.
  Training dataset size is (0.89).

- TEST SET: It is a subset used for testing of the trained model.
  Test dataset size is (0.11).

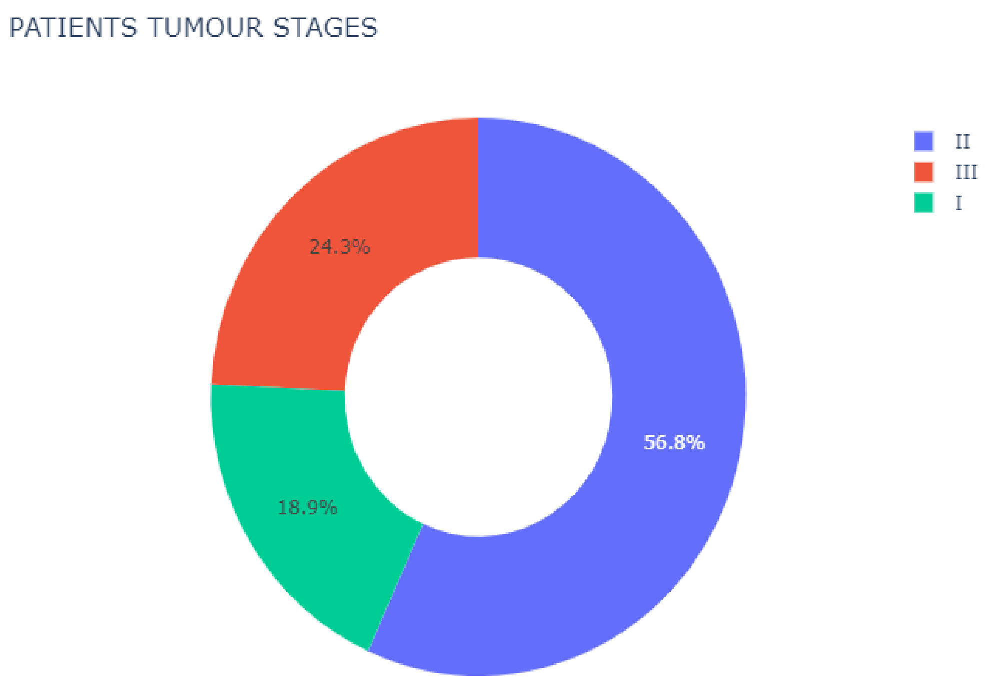
OBSERVATIONS AND RESULT

CHAPTER 6:

A) **6.1** STAGES OF TUMOUR:

◎ **6.1.1** OBSERVATION:

Chart showing the distribution of mammary carcinoma patients stages of tumour:-



The outline made it clean the percentage of patients suffering from different stages of tumour. It helps machine to learn about the ratio of patients suffering from specific stage of tumour.

◎ **6.1.2** RESULT:

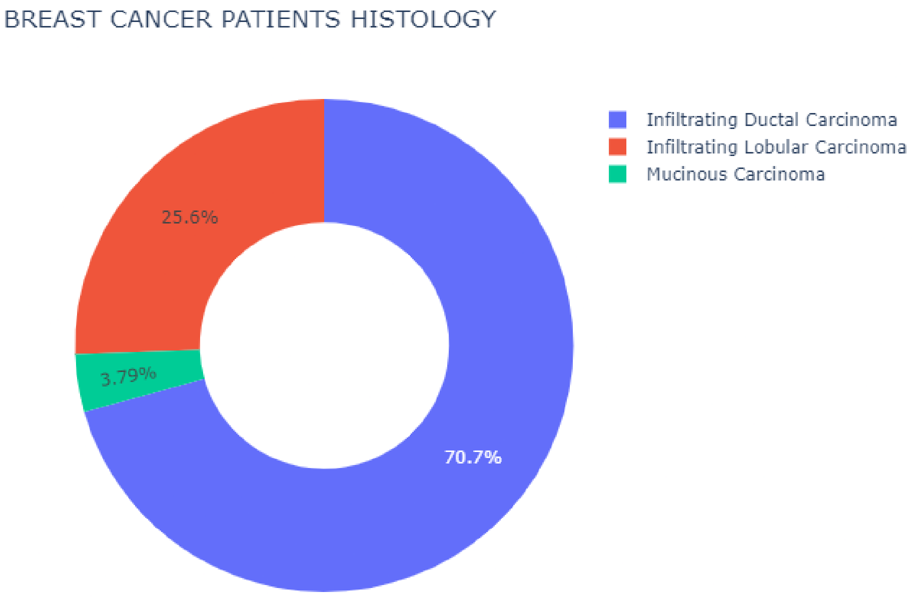
TUMOUR STAGE	PERCENTAGE OF PATIENTS
Stage I	18.9%
Stage II	56.8%

Stage III	24.3%

B) **6.2** HISTOLOGY OF CANCER PATIENTS:

◎ **6.2 .1** OBSERVATION:

Chart showing the distribution of histology of breast cancer patients: -



It is observed that more portion of the breast cancer patients are having the histology of Infilterating Ductal Carcinoma.

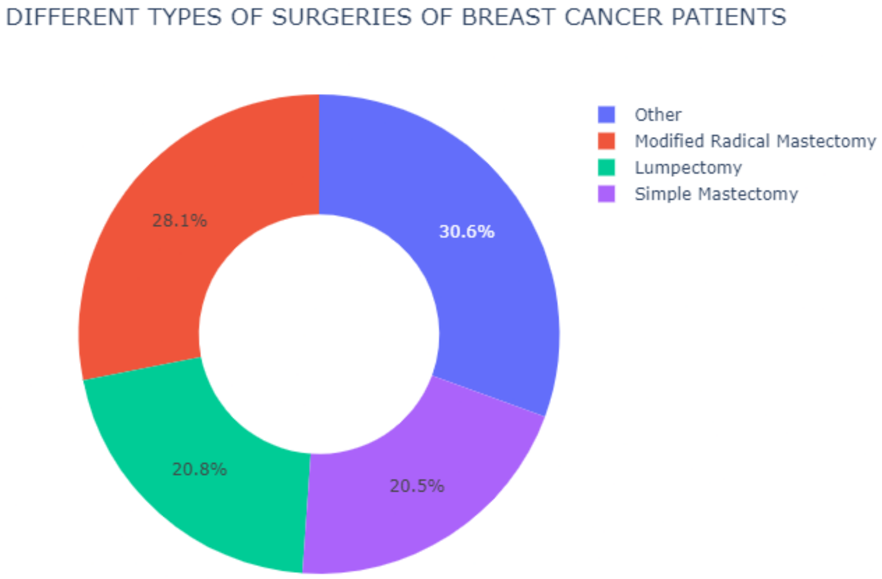
◎ **6.2 .2** RESULT:

HISTOLOGY	PERCENTAGE OF PATIENTS
Infiltrating Ductal Carcinoma	70.7%
Infiltrating Lobular Carcinoma	25.6%
Mucinous Carcinoma	3.79%

C) **6.3** TYPES OF SURGERIES OF CANCER PATIENTS:

◎ **6.3 .1** OBSERVATION:

Pie chart showing the distribution of different types of surgeries breast cancer patients underwent:



This chart will help model to learn about the proportion of Malignant tumor of breast victims underwent different surgeries and their survival rates in accordance with the type of suregery they underwent.

◎ **6.3.2** RESULT:

TYPES OF SURGERY	PERCENTAGE OF PATIENTS
Modified Radical Mastectomy	28.1%
Lumpectomy	20.8%
Simple Mastectomy	20.5%
Others	30.6%

D) **6.4** FINAL PREDICTION:

```
['Alive']
```

CONCLUSION AND FUTURE WORK

CHAPTER 7:

- ### 7.1 CONCLUSION:

The Breast Cancer Survival Prophecy framework using ML in python language is successfully trained using logistic regression model under supervised learning. This model also helped to know the real life working of regression algorithm of machine learning. Google Colab Notebook is used for implementing this model.

This model will predict whether a patient suffering from the dangerous disease like breast cancer will survive it or not post surgery by taking into account various features based on the dataset provided to it during its training.

This project also helped in knowing the working of training and testing dataset and knowledge of training of a model with machine learning techniques.

- ### 7.2 FUTURE WORK:

The assessment of the final result implies that use of dataset using machine learning in healthcare to train various models can provide propitious tools for inference in this sphere. Forthcoming exploration job ought to be carried out in this field so that this model or more relative models can predict on more parameters or more variables fluently.

More dataset should be collected to train models more efficiently. More number of ML algos should be studied that can forecast the survival rates and comparisions can be done on basis apperenting to those studies. The final outcome should be reducing error rates by providing maximum accuracy.

REFERENCES

1) For Dataset used :

https://www.kaggle.com/datasets/amandam1/breastcancerdataset?resource=download

2) For Research articles:

- https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article
- https://dl.acm.org/doi/10.1145/3492547.3492590
- https://www.javatpoint.com/supervised-machine-learning
- https://bmcmedinformdecismak.biomedcentral.com/track/pdf/10.1186/s12911-019-0801-4.pdf
- https://www.mayoclinic.org/breast-cancer/expert-answers/faq-20058066#:~:text=HER2%2Dpositive%20breast%20cancer%20is,that%20makes%20the%20HER2%20protein.
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1847991/

3) For Book:

Sheetal Taneja and Naveen Kumar, “Python Programming : A Modular Approach”, 12th impression, Pearson Publications, 2022, 17:483-496.