Research Paper on

“Breast Cancer Classification”

Department of Information Technology

2022-2023

P.C. Vashist

(Head Of Department)

Arjun Singh

(Assistant Professor)

Lakshita Singh(2001920310086)
Yashi Tripathi(2001920130197)

Department of Information Technology

G. L. Bajaj Institute of Technology and Management

Knowledge Park-3, Plot No 2, Greater Noida - 201306

GL BAJAJ
GREATER NOIDA
GL BAJAJ
Institute of Technology & Management
Department of Information Technology

Breast Cancer Classification Using Machine Learning

Lakshita Singh

Greater Noida, India 201306: Information Technology Department, G.L. Bajaj Institute of Technology & Management

Yashi Tripathi

Greater Noida, India 201306: Information Technology Department, G.L. Bajaj Institute of Technology & Management

ABSTRACT:

One of the most widespread health issues in the globe is breast cancer. Around 1.2 million more instances of breast cancer are identified every year, and 400 000 women pass away as a result.It is now the need of the world to develop efficient ways to classify the type of cancer easily. This paper proposed now presents the use and implementation of Machine Learning to classify Benign and Malignant tumors. Machine learning methods are being used to the sklearn dataset for breast cancer (data = load breast cancer()).To categorise cancer as benign or malignant, the best ideas will be employed.

About 8% of women are diagnosed with breast cancer (BC) over their lifetime, second only to lung cancer. Both in developed and underdeveloped nations, BC is the second most common cause of mortality.

Gene mutation, persistent discomfort, changes in the size, colour (redness), and skin texture of the breasts are its defining traits.

The most common categorization for breast cancer is binary (benign cancer/malign cancer), which helps pathologists discover a systematic and objective prognosis.

In this machine learning project, we will study and categorise breast cancer (which group it falls under), as there are primarily two types of breast cancer:

-> Breast cancer of the malignant

-> Benign variety

KEYWORDS:

Classification of benign and malignant breast cancer , Logistic Pre-Processing, Train, Test, Split, Datapre-processing , Breast cancer dataset, Numpy , Pandas, model selection.

Introduction:

The spread of abnormal cell growth into the body's surrounding tissues leads to the formation of the cancerous tumour. The lack of a breast tumour is seen to be normal. Benign or malignant tumours can be distinguished.The benign tumour cells cannot invade surrounding tissue and can only grow locally Malignant tumours are harmful cells that can grow uncontrolled, invade adjacent tissue, and spread to numerous sections of the body. Breast cancer comes in a wide variety of forms with varying genetic make-up, aggressiveness, and stages of dissemination. When breast cancer is found sooner using mammography, the likelihood of survival may rise. Increased caseloads for radiologists would follow the deployment of mass screening. This will raise the likelihood of an incorrect diagnosis. The radiologist would be helped to find breast cancer by the categorization using logistic regression.

There are three basic components of a breast: connective tissue, ducts, and lobules. The glands that generate milk are called lobules. The tubes that bring milk to the nipple are called ducts. The connective tissue, which is made up of fatty and fibrous tissue, envelops and holds everything in place. Most breast cancers start in the ducts and lobules.

Through lymphatic and blood arteries, this cancer is capable of metastasizing outside the breast. Breast cancer is referred regarded as having metastasized when it spreads to other bodily areas.

What Leads To The Development Of Cancer In The Body?

The basic units of tissue, cells, are where cancer first manifests itself. The breast and other areas of the body contain tissues. Sometimes, the process of cell growth goes wrong, resulting in the formation of new cells when the body doesn't need them and the failure of old or damaged cells to die as they should. When all of these things happen, a buildup of cells frequently results in the formation of a mass of tissue known as a lump, growth, or tumour.

Just what is a tumour?

A tumour, also known as a "neoplasm," is an abnormal mass of tissue that develops when cells grow and divide much more quickly than they should or when they don't die as they should. a benign tumour is one that does not spread to other areas of the body and does not suggest malignancy. However, certain tumours are dangerous or malignant, and they can spread to other parts of the body through the lymphatic and blood systems.

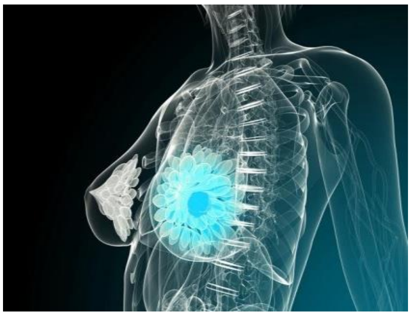


Fig: Tumor

The most common malignancy in women is breast cancer, therefore it is essential to have an accurate and trustworthy approach for determining whether a tumour is benign or malignant. Modern approaches for machine learning and Fine Needle Aspiration (FNA) cytology data allow for more accurate detection and early diagnosis of this cancer. In this research, we suggest a method to categorise tumours into benign and malignant, and logistic regression has been applied for this aim.

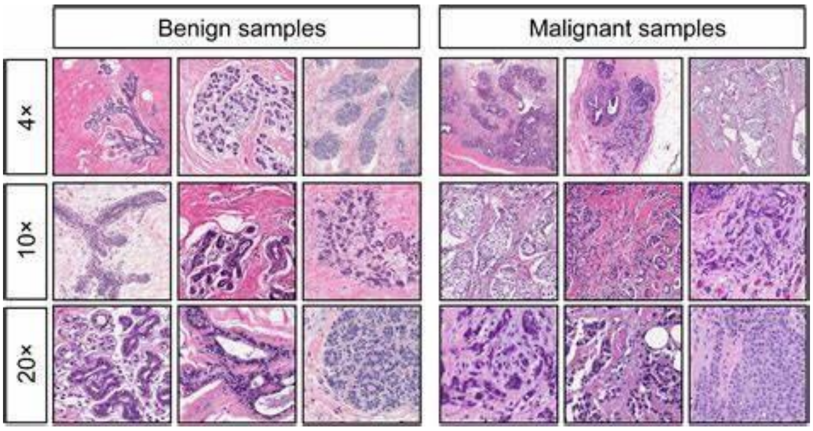


Fig: Samples of both Tumors

- EXISTING WORK

Breast cancer can now be identified through a physical examination, mammography, or a biopsy. The four methods for a biopsy are fine needle aspiration (FNA), core needle biopsy, surgical biopsy, and lymph node biopsy. As opposed to that, mammography is a frequent technique for detecting breast cancer, but only board-certified radiologists should interpret the findings. Since different radiologists get different conclusions from a single mammography, this could result in various interpretations [3]. Mammography has a 68 to 79 percent accuracy rate. When a tumour is found by mammography, a biopsy is performed to determine how cancerous it is. Although surgical biopsy is almost always accurate, it is expensive, aggressive, time-consuming, and unpleasant.

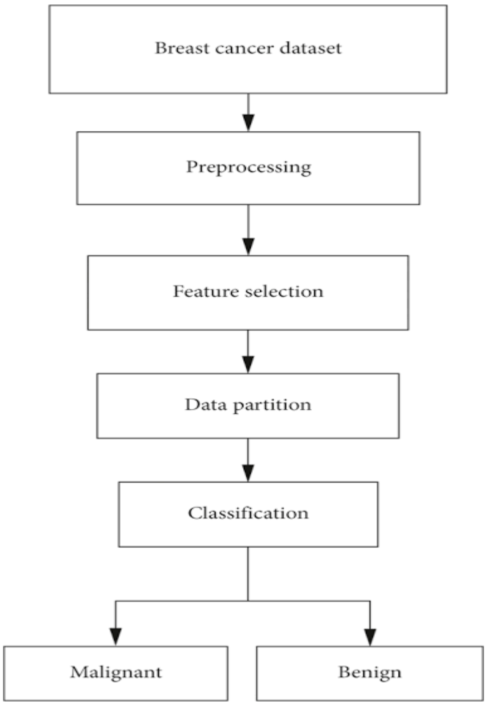
Due to its outpatient nature, cheap cost, simplicity, and speed, FNA is one of the diagnostic techniques mentioned above that is frequently used to detect breast cancer [4]. The accuracy of FNA with visual interpretation might range from 35 to 95 percent depending on the doctor's experience. [3]. This procedure involves studying the cytological characteristics of the liquid that was taken from breast tissue seen via a microscope. The cytology results can be examined to determine the tumor's malignancy [2]. The use using data mining and machine learning methods can be crucial in the identification of tumours when the doctor is uncertain and unable to establish whether they are benign or malignant.

In fact, using a computer-based method to find breast cancer in its earliest stages can lengthen the time patients survive after being diagnosed with the disease. Thus, throughout the past ten years, research has been done to identify more precise methods of diagnosing breast cancer.

Patients are typically divided into the benign class (those without cancer) and the malignant class using computer-based breast cancer procedures (which have cancer). There are many clever ways to categorise data about breast cancer. [2], [3], [4], [5], [6], [7], [8], [9], [10], [11],[12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], whereas others carry out categorization without feature selection. Some of them incorporate techniques for feature selection. For instance, Naive Bayes (NB) and Weighted Nave Bayes classifiers have been used in a number of research to identify the kind of tumour; however, in some of them, feature selection has been undertaken prior to classification, while in others, it has not In order to detect breast cancer popular classifiers likeSupport vector machine Additionally, Decision Tree (DT), Random Forest (RF), and Vector Machine (SVM) have been employed. In one study, Principal Component Analysis was used to identify features (PCA) and Relief were employed. For classification, Gaussian Bayes is also used in conjunction with Linear Discriminant Analysis (LDA). Data classification using the Mamdani fuzzy inference system has been done in another study utilising the same feature selection technique. Previous studies have also employed neural networks to forecast breast cancer. One of these attempts,features are chosen using a genetic algorithm prior to classification (GA). The application of evolutionary algorithms as feature selection techniques and classifiers, including GA and Particle Swarm Optimization (PSO), is documented in the literature. To assist doctors, researchers are working to provide better options.

- PROPOSED WORK

  As far as we are aware, no prior studies have employed the logistic regression technique, which is used in this study to weed out inefficient characteristics. This linear, efficient, and economical algorithm is used to successfully choose features. Actually, the characteristics are weighted, and by using these weights, extra features are eliminated.



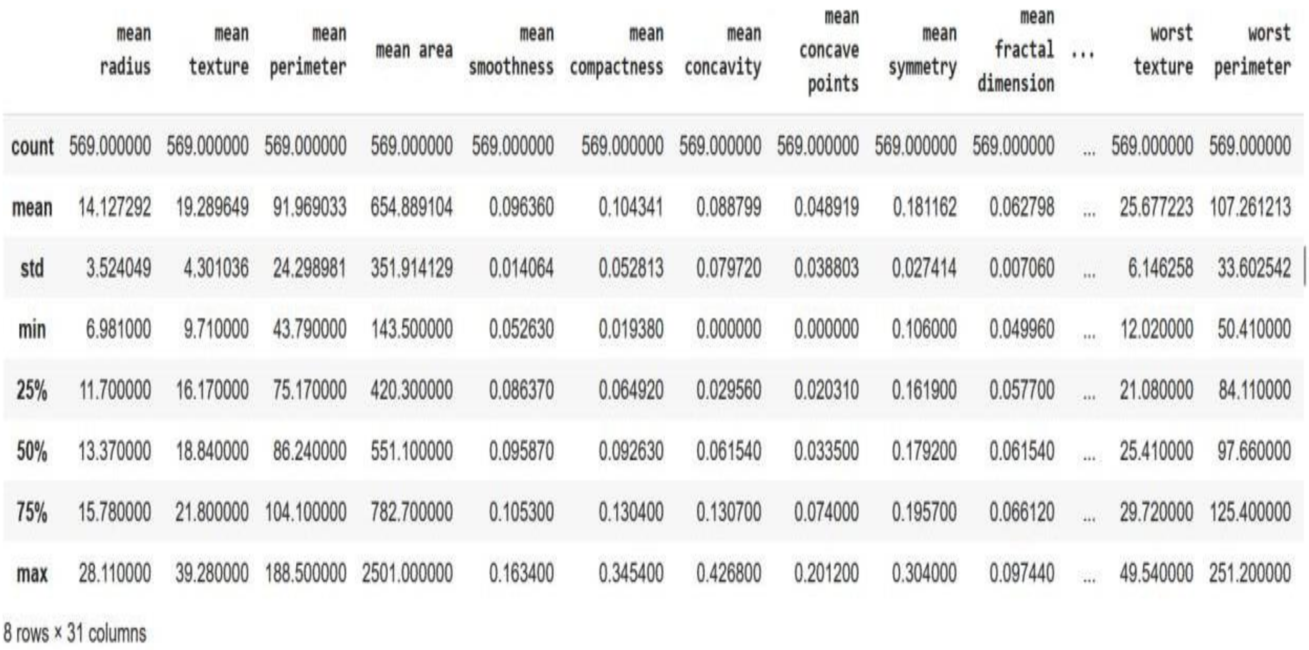
Briefly, this paper's contributions are as follows:

- Making use of various datasets to assess the suggested techniques
- achieving high accuracy for specified datasets
- lowering the computational cost by employing a fresh method of feature selection based on logistic regression

- THE DATASET

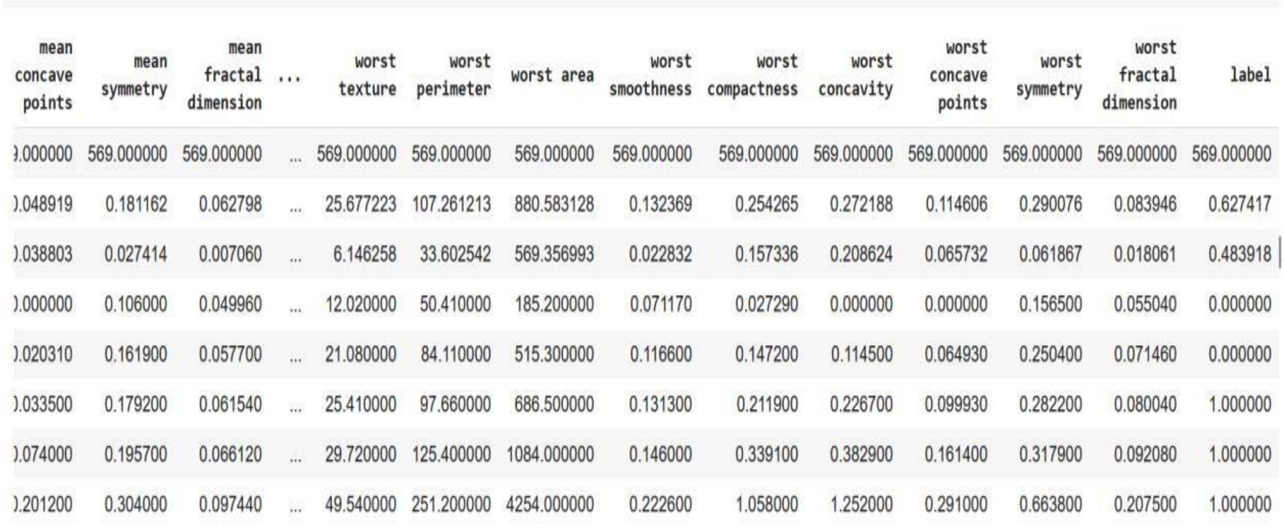
We will be using a dataset on breast cancer from Sklearn.
We will thoroughly study the dataset, which will answer all of the queries about the dataset we will use, the no. of columns and rows etc. Consequently, there are 30 columns that are:

- Mean Radius
- Mean Texture
- Mean Perimeter
- Mean Area
- Mean Smoothness
- Mean Compactness
- Mean Concavity
- Mean Concave Points
- Mean Symmetry
- Mean Fractal Dimension
- Radius Error
- Texture Error
- Perimeter Error
- Area Error
- Smoothness Error
- Compactness Error
- Concavity Error
- Concave Points Error
- Symmetry Error
- Fractal Dimension Error
- Worst Radius
- Worst Texture
- Worst Perimeter
- Worst Area
- Worst Smoothness
- Worst Compactness
- Worst Concavity
- Worst Concave Points
- Worst Symmetry
- Worst Fractal Dimension
- Label



	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	...	569.000000	569.000000
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.181162	0.062798	...	25.677223	107.261213
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.027414	0.007060	...	6.146258	33.602542
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000	0.049960	...	12.020000	50.410000
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900	0.057700	...	21.080000	84.110000
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.179200	0.061540	...	25.410000	97.660000
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700	0.066120	...	29.720000	125.400000
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.304000	0.097440	...	49.540000	251.200000

8 rows × 31 columns



mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension	label
).000000	569.000000	569.000000	...	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
).048919	0.181162	0.062798	...	25.677223	107.261213	880.583128	0.132369	0.254265	0.272188	0.114606	0.290076	0.083946	0.627417
).038803	0.027414	0.007060	...	6.146258	33.602542	569.356993	0.022832	0.157336	0.208624	0.065732	0.061867	0.018061	0.483918
).000000	0.106000	0.049960	...	12.020000	50.410000	185.200000	0.071170	0.027290	0.000000	0.000000	0.156500	0.055040	0.000000
).020310	0.161900	0.057700	...	21.080000	84.110000	515.300000	0.116600	0.147200	0.114500	0.064930	0.250400	0.071460	0.000000
).033500	0.179200	0.061540	...	25.410000	97.660000	686.500000	0.131300	0.211900	0.226700	0.099930	0.282200	0.080040	1.000000
).074000	0.195700	0.066120	...	29.720000	125.400000	1084.000000	0.146000	0.339100	0.382900	0.161400	0.317900	0.092080	1.000000
).201200	0.304000	0.097440	...	49.540000	251.200000	4254.000000	0.222600	1.058000	1.252000	0.291000	0.663800	0.207500	1.000000

Fig:The Dataset

WORKFLOW :

Following is the process for classifying breast cancer using machine learning:

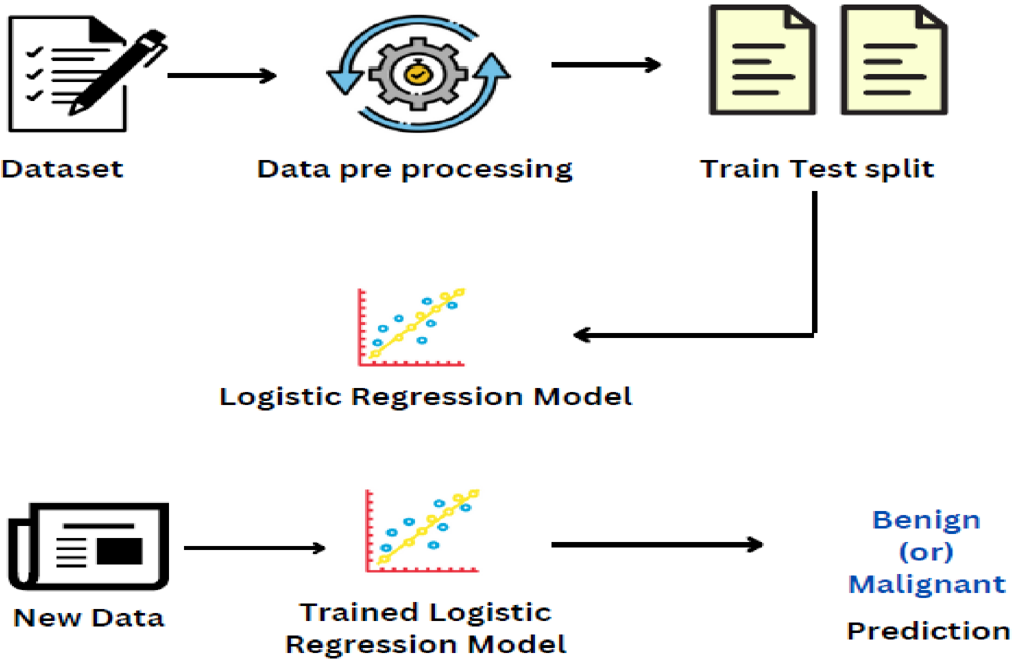


Fig: Workflow

- **DATASET**
  - A type of biopsy procedure known as fine needle aspiration involves inserting a thin needle into the area of tissue that appears abnormal. This technique can be used to diagnose conditions like cancer.
  - Breast Cancer Dataset is taken from sk.learn datasets.
  - *breast_cancer_dataset = sklearn.datasets.load_breast_cancer()*

- **PRE-PROCESSING**
  - Data must be pre-processed in order to be training data and testing model are separated.
  - Train the model of machine learning using train and & evaluate, the model using the Test data.
  - To better comprehend and interact with the data, we will now transform it to a dataframe using pandas. To understand what the data contains, print the dataset's first five samples using the head function.
  - Data can be split into testing and training data using the Train Test Split tool.

- **LOGISTIC REGRESSION MODEL:**
  - Train the model of machine learning in model of Logistic Regression , as logistic regression model uses Binary Classification.
  - Binary Classification to tell whether the data is Benign or Malignant.
- **EVALUATION OF MODEL :**
  - It is done using accuracy score() ,where it will give the reliability of training and test data.
  - The new data is classified into Benign or Malignant.
  - The outcome indicates whether the input tumour is a benign or malignant tumour.

RESULT :

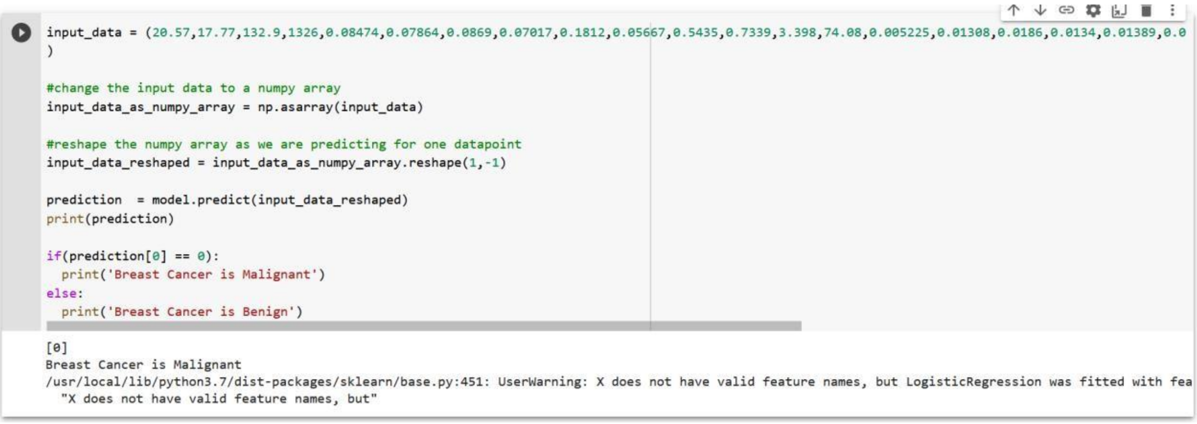
The trained data is evaluated and therefore classified in Benign and Malignant.

➔ 0 represents Malignant

➔ 1 represents Benign

Evaluation is done using the accuracy score() which tells the accuracy of the classification.

```
➔ Accuracy on training data =  0.9472527472527472
➔ Accuracy on test data =  0.9210526315789473
```



```
input_data = (20.57,17.77,132.9,1326,0.08474,0.07864,0.0869,0.07017,0.1812,0.05667,0.5435,0.7339,3.398,74.08,0.005225,0.01308,0.0186,0.0134,0.01389,0.0
)

#change the input data to a numpy array
input_data_as_numpy_array = np.asarray(input_data)

#reshape the numpy array as we are predicting for one datapoint
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction  = model.predict(input_data_reshaped)
print(prediction)

if(prediction[0] == 0):
  print('Breast Cancer is Malignant')
else:
  print('Breast Cancer is Benign')
```

```
[0]
Breast Cancer is Malignant
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:451: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with fea
  "X does not have valid feature names, but"
```

The above image clearly tells us that the input data is a case of malignant tumor, which is cancerous.

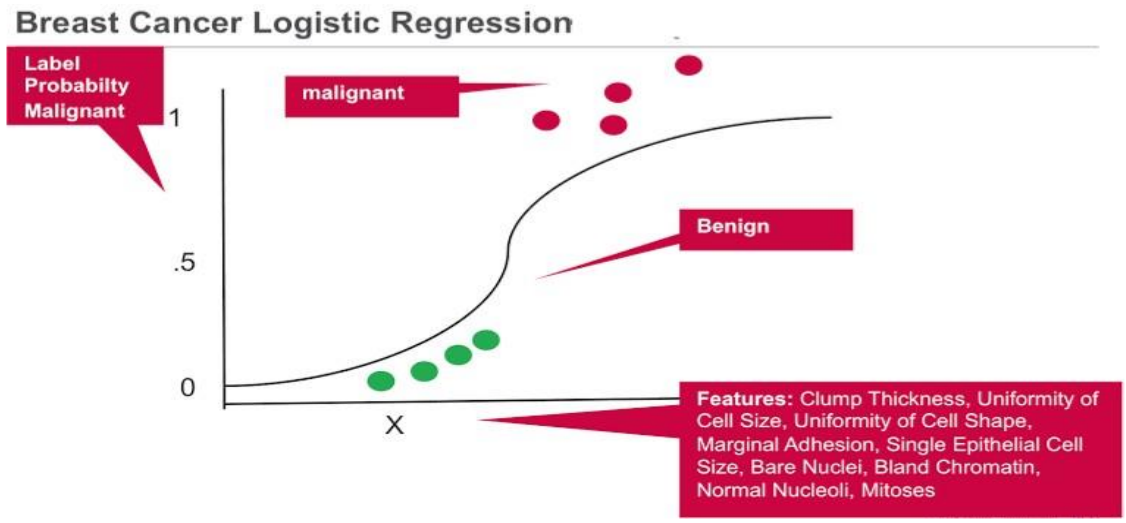

Breast Cancer Logistic Regression
Label Probabilty Malignant
1
.5
0
malignant
Benign
X
Features: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses

Fig: depicting how Logistic Regression Model is used for binary classification.

Table 1 and the figure below show the accuracy percentage for Breast Cancer Diagnostics datasets. From the results of training set and testing set, we infer that various classifiers have varying accuracies, out of which we in have been working upon the *Logistic Regression Classifier.*

Algorithms	Accuracy Training Set (%)	Accuracy Testing Set (%)
SVM	98.4%	97.2%
Random Forest	99.8%	96.5
Decision Tree	98.8%	95.1%
Logistic Regression	95.5%	95.8%
KNN	94.6%	93.7%

Table 1: Accuracy percentage for breast cancer diagnostic dataset.

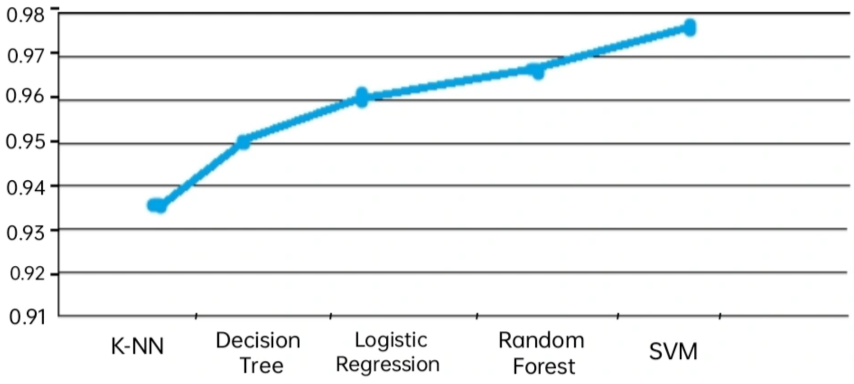


Figure: Comparative graph of different classifiers.

CONCLUSION AND NEXT WORK :

The best collection of characteristics that predict breast cancer were selected for this study using the logistic regression approach.

Using a logistic regression model, we were able to create a method for categorising breast cancer that is fairly accurate and useful for recognising the main types of breast cancer.Major health risks can be avoided and even lives may be saved through the early detection and treatment of various conditions.

Our model has a 94.8% training data accuracy and a 92.9% testing data accuracy. Our model is identifying test data quite effectively and precisely, as we have already witnessed.As a result, using the Numpy and pandas libraries, we have learned how to analyse and visualise data.

One needs to be familiar with machine learning principles and the Python language to make this project work.

This project effectively distinguishes which one is benign and which is malignant breast cancer.
To obtain an accurate result, a total of 30 columns from the obtained dataset were used. The algorithm classifies the data into benign and malignant once the characteristics have been retrieved and fully trained.

In the future, we will be able to provide a picture as an input instead of numerical data, and the classifier will then have to determine if the image is benign or cancerous before classifying it.

Future enhancements to the system might include other capabilities like the suggestion of drugs or therapies based on the severity of the illness. Doctors can more effectively detect and treat the condition with the aid of this prediction and suggestion system.

REFERENCES:

1.

[2] R. Sheikhpour, M. Agha Sarram, R. Sheikhpour

Using classifiers based on kernel density estimation for the detection of breast cancer and using particle swarm optimization to choose the best features

Appl. Soft Comput., 40 (2016), pp. 113-131

[3] M.A. Karabatak

A new Naive Bayesian classifier for the identification of breast

cancer

Measurement, 32–36 (2015)

Google Scholar

[4] S. Eslami, H. Shahraki, Sh. Aalaei, and A.R. Rowhanimanesh

Three separate datasets are used in an attempt to pick breast cancer-related features using a genetic algorithm.
Iranian Journal of Basic Medical Sciences, 19, 476-482 (2016)

[5] C.P.Sumathi and B.M. Gayathri

A computerised Method for Classifying Breast Cancer Using the Gaussian Naive Bayes Classifier

International Journal of Computer Application, 148 (6) (2016)

[6] B. M. Gayathri, C. P. Sumathi, "Mamdani Fuzzy Inference system for Breast cancer risk identification," IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai, India, December 2015.

[7] "Fuzzy technique for pre- detection of breast cancer from the Fine Needle Aspirate analysis," Biomedical Engineering Online, 11(83), 2012. G. R. Sizilio, C. R. Leite, A. M. Guerreiro, and A. D. D. Neto

[8] O. Ibrahim, M. Shahmoradi, H. Ahmadi, and M. Nilshahi

A fuzzy logic-based knowledge-based approach for classifying breast cancer

Telematics Education., 34 (2017), pp. 133-144

[9] S. Kumar Mandal, A. Gupta, and A. Hazra, "International Journal of Computer Applications," "Study and analysis of breast cancer cell identification using naive Bayes, SVM, and ensemble approaches," 145(2), 39-45, July 2016.

[10] N. Modi and K. Ghanchi, "Comparative study of feature selection strategies and associated machine learning algorithms (WBCD)", International Conference on ICT for Sustainable Development, 1 (2016), pp. 215-224.

[11] Building Minimal Classification Rules for Breast Cancer Diagnosis, 10th International Conference on Knowledge and Smart Technology (KST), Chiang Mai, Thailand, Jan.–Feb. 2018. Ph. Douangnoulack, V. Boonjing.

[12] A. Subasi , E. Alickovic,

utilising Rotation Forest Neural Computing and GA feature selection to

diagnose breast cancer.Appl., 28 (2015), pp. 753-763

[13] Z. Hussain, M.K. Osman, S.N. Sulaiman, F. Ahmad, N.A. Mat Isa, and Pattern Analysis and Applications, Nov., 18 (4) (2015), pp. 861-870. Breast cancer detection utilising an ANN with feature selection and parameter optimization based on GA.

[14] An Effective Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets, International Journal of Advanced Research in Computer Science and Software Engineering, 272-277, 2014. G. Ravi Kumar, G. A. Ramachandra

[15] P. Afshar, A. Ahmadi

Intelligently detect breast cancer using support vector machines and particle swarm optimization.

J. Exp. Theor. Artif. Intell., 28 (6) (2016), pp. 1021-1034

[16] E. Bilgili, A. Akan, A. Mert, N. Kılıç

Breast Cancer Detection Using a Reduced Feature Set Using Computer Mathematics in Medicine (2015)

[17] V. Singh , M. Kumari

Breast Cancer Prediction system

Procedia Computer Science, 132 (2018), pp. 371-376

[18] R. Tina, S. Sherekar , S. Patil

International Journal Of Computer Science And Applications, 6 (2) Performance Evaluation of the J48 Classification Algorithm and Naive Bayes for Data Classification (2013)

[19] T. Noel, H. Asri, H. Al Moatassime, and H. Mousannif

Detecting and predicting breast cancer risk using machine learning techniques Procedia Computer Science (2016)

[20] A comparison of machine learning techniques for detecting and diagnosing breast cancer by D. Bazazeh and R.M. Shubair

5th International Election Conference

[21] "Hybrid multistage fuzzy clustering system for medical data categorization," International Conference on Computing Sciences and Engineering (ICCSE), Kuwait City, Kuwait, March 2018. M. S. Abdullah, F. AL-Anzi, and S. Al-Sharhan.

[22] L. Sahinbegoviü and A. Abdel-Ilah, "Using machine learning technique in categorization of breast cancer," Springer Nature Singapore Pte Ltd, 3-8, 2017.

[23] A. Bhardwaj and A. Tiwari, "Genetically optimised neural network model for breast cancer detection," Expert Syst (2015)

[24] F. Li, J. Yang, L. Peng, W. Chen, W. Zhou, and J. Zhang

A semi-supervised system for the detection of breast cancer that is immune-inspired

259–265 in Comput. Methods Programs Biomed., vol. 134 (2016).

[25] Using multi-classifiers, G.I. Salama, M.B. Abdelhalim, and M. Abd-elghany Zeid diagnosed breast cancer on three different datasets. 1 (1) (2012), pp. 36–43, International Journal of Computer and Information Technology

[26] "Ensemble-Based Hybrid Approach for Breast Cancer Data," International Conference on Communications and Cyber Physical Engineering, 713–720, 2018. G. Naga RamaDevi, K. Usha Rani, and D. Lavanya

[27] Enhanced Generalized Regression Neural Net for Breast Cancer Detection, S. Babaei Ghalejoughi and N. Lotfivand, International Journal of Computer Science and Information Security, 16(1), 2018.