BREAST CANCER PREDICTION MODEL USING MACHINE LEARNING

Summary Sheet submitted in fulfillment of sem-5 requirements for degree of

BACHELOR OF TECHNOLOGY

By

ADITI JAIN - 20103273

NANDINI AGRAWAL - 20103144

ISHIKA JAIN - 20103267

Department of Computer Science and Technology

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY
(Declared Deemed to be University U/S 3 of UGC Act)
A-10, SECTOR-62, NOIDA, UTTAR PRADESH, INDIA

Under supervision of

DR. SHIKHA JAIN

November 2022

ACKNOWLEDGEMENT

We would like to place on record my deep sense of gratitude to **Dr. Shikha Jain**, **Assistant Professor**, Jaypee Institute of Information Technology, India for her generous guidance, help and useful suggestions.

We express our sincere gratitude to **Dr. Vikas Saxena (HOD), Dept. of CSE/IT** Jaypee Institute of Information Technology, India, for his stimulating guidance, continuous encouragement, and supervision throughout the course of present work.

We also wish to extend my thanks to my friends, team members and other classmates for their insightful comments and constructive suggestions to improve the quality of this project work.

Signature of Students

NANDINI AGRAWAL (20103144)

ISHIKA JAIN (20103267)

ADITI JAIN (20103273)

-

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and beliefs, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma from a university or other institute of higher learning, except where due acknowledgment has been made in the text.

Place: Noida

Date: 04-12-2022

Name: Nandini Agrawal

Enrolment No.: 20103144

Name: Ishika Jain

Enrolment No.: 20103267

Name: Aditi Jain

Enrolment No.: 20103273

CERTIFICATE

This is to certify that the work titled "**BREAST CANCER PREDICTION MODEL USING MACHINE LEARNING** " submitted by **Nandini Agrawal, Ishika Jain, Aditi Jain** in partial fulfillment for the award of degree of B. Tech of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of any other degree or diploma.

Dr. Shikha Jain

04-12-2022

LIST OF FIGURES

ABSTRACT

In 2020, there were 2.3 million women diagnosed with breast cancer and 685 000 deaths globally. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world's most prevalent cancer.

Factors increase the risk of breast cancer including increasing age, obesity, harmful use of alcohol, family history of breast cancer, history of radiation exposure, reproductive history (such as age that menstrual periods began and age at first pregnancy), tobacco use and postmenopausal hormone therapy.

Breast cancer treatment can be highly effective, achieving survival probabilities of 90% or higher, particularly when the disease is identified early.

Machine Learning (ML) allows us to discover relations between prognostic factors and to predict breast cancer prognosis. These models might become an additional resource in our daily clinical practice.

In this project, we present a predictive model to analyze breast cancer using Machine Learning.

For this purpose we analyzed several Machine Learning Algorithms which will help specialists in diagnosing whether the features of a person's medical data point towards his cells being malignant or benign, which further helps in decision making in treatment method for the same purpose.

TABLE OF CONTENTS

CHAPTER-1
INTRODUCTION

Breast cancer is a disease in which cells in the breast grow out of control. There are different kinds of breast cancer. The kind of breast cancer depends on which cells in the breast turn into cancer. With being the most common type of cancer in women, breast cancer accounts for 14% of cancers in Indian women. It is reported that with every four minutes, an Indian woman is diagnosed with breast cancer. Breast cancer is on the rise, both in rural and urban India.

Problem Statement:-

In this project, we present a predictive model to analyze breast cancer. For this purpose we analyzed several Machine Learning Algorithms which will help specialists in diagnosing whether the features of a person's medical data point towards his cells being malignant or benign, which further helps in decision making in treatment method for the same purpose.

Motivation behind the project:-

As per the Globocan data 2020, in India, breast cancer accounted for 13.5% of all cancer cases and 10.6% (90408) of all deaths. One in twenty-eight Indian women is likely to develop breast cancer during her lifetime.

Many of the breast cancer patients stay undetected until it's very late. Doctors are not able to catch the early signs of onset of breast cancer, as a result several patients lose their lives. Building a robust machine learning model can effectively reduce the number of undetected cases and help in early detection of breast cancer when it is most treatable.

Classification of data as malignant or benign:-

In this project, the data points have been classified as Malignant or Benign using classifiers:

Classification is the process of predicting the class of given data points. Classes are sometimes called targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

For instance our problem statement of Breast Cancer Prediction can be classified as a classification problem. This is binary classification since there are only 2 classes "Malignant" and "Benign". A classifier utilizes some training data to understand how given input variables relate to the class. In this case, known malignant denoted by a "0" and benign denoted by a "1" emails have to be used as training data. When the classifier is trained accurately, it can be used to detect breast cancer.

Classification belongs to the category of supervised learning where the targets are also provided with the input data. There are many applications in classification in many domains such as in credit approval, medical diagnosis, target marketing etc.

In this project, three algorithms: Logistic Regression, KNN and Naive Bayes have been implemented in order to achieve the desired classification of data into malignant or benign.

CHAPTER 2

LITERARY SURVEY

This is a literature review on the practice of Breast Cancer Prediction using machine learning algorithms which explores a part of the current debates about the subject. It was conducted using an explorative and unstructured method of review which explored breast cancer prediction projects by studying different articles and papers that offered insights and perspective and discipline. The purpose of the review is to identify different perspectives that exist on the practice of breast cancer prediction.

Breast cancer is one of the major diseases that cause a high number of women's deaths. To decrease these numbers, early diagnosis is an important task in the medical process. Machine learning techniques are an effective way to classify data especially in the medical field, where those methods are widely used in diagnosis and decision making.

The healthcare environment is one of the most accurate fields for data science applications due to the amount of data that it contains and the suitability of data type. The flow of data in hospitals is a continuous process and includes numerical values in general. Healthcare is an open system for improvements with studies about data mining and machine learning techniques.

One of the models built by Dr. Megha Rathi and Dr. Arun Kumar Singh on breast cancer detection using machine learning algorithms is discussed here. The work in this paper is focusing on Breast cancer prediction using Naive bayes. Algorithms by using Java Netbeans interface then compared the result in other algorithms using WEKA. The method in this paper includes the screening of dataset from UCI ML repository and it consists of 699 instances and 10 attributes. It has positive samples and negative samples and every sample has 10 attributes defined for them.

The method in this paper includes the segments of Background study, Methodology and

Experimental Study of the clinical data and the outcome.The conclusion and future work of this paper was to assist oncologists in decision making for breast cancer patients by stating that their approach performs better and provides better accuracy in predicting the cancer types as benign and malignant. Under methodology it is stated that from the confusion matrix to analyze the performance criterion for the classifiers in detecting breast cancer, accuracy, precision (for multiclass dataset), sensitivity and specificity have been computed to give a deeper insight of the automatic diagnosis.

Another research on breast cancer prediction using machine learning algorithms uses comparison with BCRAT and BOADICEA here under the methodology the have quantified and compared the performance of eight different ML methods to the performance of BCRAT and BOADICEA using eight simulated datasets and two retrospective samples: a random population-based sample of U.S. breast cancer patients and their cancer-free female relatives (*N* = 1143), and a clinical sample of Swiss breast cancer patients and cancer-free women seeking genetic evaluation and/or testing (*N* = 2481). The result of this study implied that Predictive accuracy (AU-ROC curve) reached 88.28% using ML-Adaptive Boosting and 88.89% using ML-random forest versus 62.40% with BCRAT for the U.S. population-based sample. Predictive accuracy reached 90.17% using ML-adaptive boosting and 89.32% using ML-Markov chain Monte Carlo generalized linear mixed model versus 59.31% with BOADICEA for the Swiss clinic-based sample. Followed by conclusion that there was a striking improvement in the accuracy of classification of women with and without breast cancer achieved with ML algorithms compared to the state-of-the-art model-based approaches. High-accuracy prediction techniques are important in personalized medicine because they facilitate stratification of prevention strategies and individualized clinical management.

Another method was proposed using KNN algorithm with an objective to develop a sophisticated and automated diagnostic system that yields accurate and reproducible results for predicting whether a breast cancer tumor is benign (non-cancerous) or malignant (cancerous). The study has implemented KNearest Neighbour Algorithm using various normalization techniques and distance functions at different values of K. A comparative study using various normalization techniques, i.e., Min-Max normalization, Z-Score normalization and Decimal Scaling normalization, and different distance metrics, i.e., Manhattan distance, Euclidean distance, Chebyshev distance and Cosine distance has been done. The accuracy of each variation is tested and the maximum accurate prediction is considered for the result. Highest accuracy of 98.24% is achieved, with KNN implementation using Manhattan distance metric, at K=14, along with Decimal scale normalization.

Another research for building machine learning model compatible with breast cancer prediction was proposed by S. Murugan, B. Muthu Kumar and S. Amudha by building a model that uses Data to analyze and predict breast cancer obtained from UCI Machine Learning Repository (Wisconsin Breast Cancer). The main objective is to classify whether the type of cancer is benign or malignant. Based on the available data set and the patient record, whether the disease is curable or incurable is predicted. Thus the success rate of classification is 84.14% and the prediction percentage is 88.14%.

One of the experimental methodology using image processing was proposed by Moh'd Hadidi, Abdulsalam Alarabeyyat and Mohannad Alhanahnah The proposed method for breast cancer detection consists of two main parts: image processing techniques and the machine learning algorithms where applying these algorithms were done by using Matlab software. In this work we extracted 209 images for 50 patient cases who have breast cancer and the testing stage was applied on many people either they have breast cancer or they have not. In this work, a supervised learning algorithm has been used. Indeed, we used two types of supervised machine learning algorithms which were the Logistic Regression and the Backpropagation neural Network and results were compared from both of them.

[1] Megha Rathi, Arun Kumar Singh, Breast Cancer Prediction using Naïve Bayes Classifier

[2] Chang Ming, Valeria Viassolo, Nicole Probst-Hensch, Pierre O. Chappuis, Ivo D. Dinov & Maria C. Katapodi, Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models

[3] Shagun Chawla, Rajat Kumar, Ekansh Aggarwal, Sarthak Swain, Breast Cancer Detection Using K-Nearest Neighbor Algorithm

[4] K. Devasena and J. Shana 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1166 012029, Building Machine Learning Model for Predicting Breast Cancer Using Different Regression Techniques.

[5] S. Murugan; B. Muthu Kumar; S. Amudha, Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest

[6] Fatima Noreen, Li Liu, Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis

[7] Nikita Rane, Jean Sunny, Rucha Kanade, Prof. Sulochana Devi, Breast Cancer Classification and Prediction using Machine Learning

[8] Mohammed Amine Najia, Sanaa El, Filalib Kawtar, Aarikac, EL Habib Benlahmard, Rachida Ait Abdelouhahide, Olivier Debauchef, Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis

[9] Shanjida Khan, Romana Rahman Ema, Cancer Disease Prediction Using Naive Bayes, K-Nearest Neighbor and J48 algorithm

[10] Sulistiani, Windu Wulandari, Fathia Dwi Astuti, Widodo Breast Cancer Prediction Using Random Forest and Gaussian Naïve Bayes Algorithms

[11] M. Vijay Anand, B. KiranBala, R. Srividhya, Kavitha C., Mohammed Younus and Md Habibur Rahman, Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer

[12] Shika Agrawal, Shweta Kharya, Sunita Soni, Naive Bayes Classifiers: A Probabilistic Detection Model for Breast Cancer

[13] Tawseef Ayoub Shaikh, Rashid Ali, A CAD Tool for Breast Cancer Prediction using Naive Bayes Classifier

[14] Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S, Breast Cancer Prediction using Machine Learning

[15] Moh'd Hadidi, Abdulsalam Alarabeyyat, Mohannad Alhanahnah, Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm

CHAPTER 3

REQUIREMENT ANALYSIS

This project can run on commodity hardware. We ran the entire project on an Intel I5 processor with 8 GB Ram, 2 GB Nvidia Graphic Processor, it also has 2 cores which run at 1.7 GHz, 2.1 GHz respectively. First part is the training phase which takes 10-15 mins of time and the second part is the testing part which only takes a few seconds to make predictions and calculate accuracy.

3.1 HARDWARE REQUIREMENT

- RAM: 4 GB
- Storage: 500 GB
- CPU: 2 GHz or faster
- Architecture: 32-bit or 64-bit

3.2 SOFTWARE REQUIREMENT

- Python 3.5 in Google Colab is used for data pre-processing, model training and prediction.
- Operating System: Windows 7 and above or Linux based OS or MAC OS

CHAPTER - 4
PROPOSED MODEL

4.1. Design of complete project model

4.1.1 DETAILED DESIGN

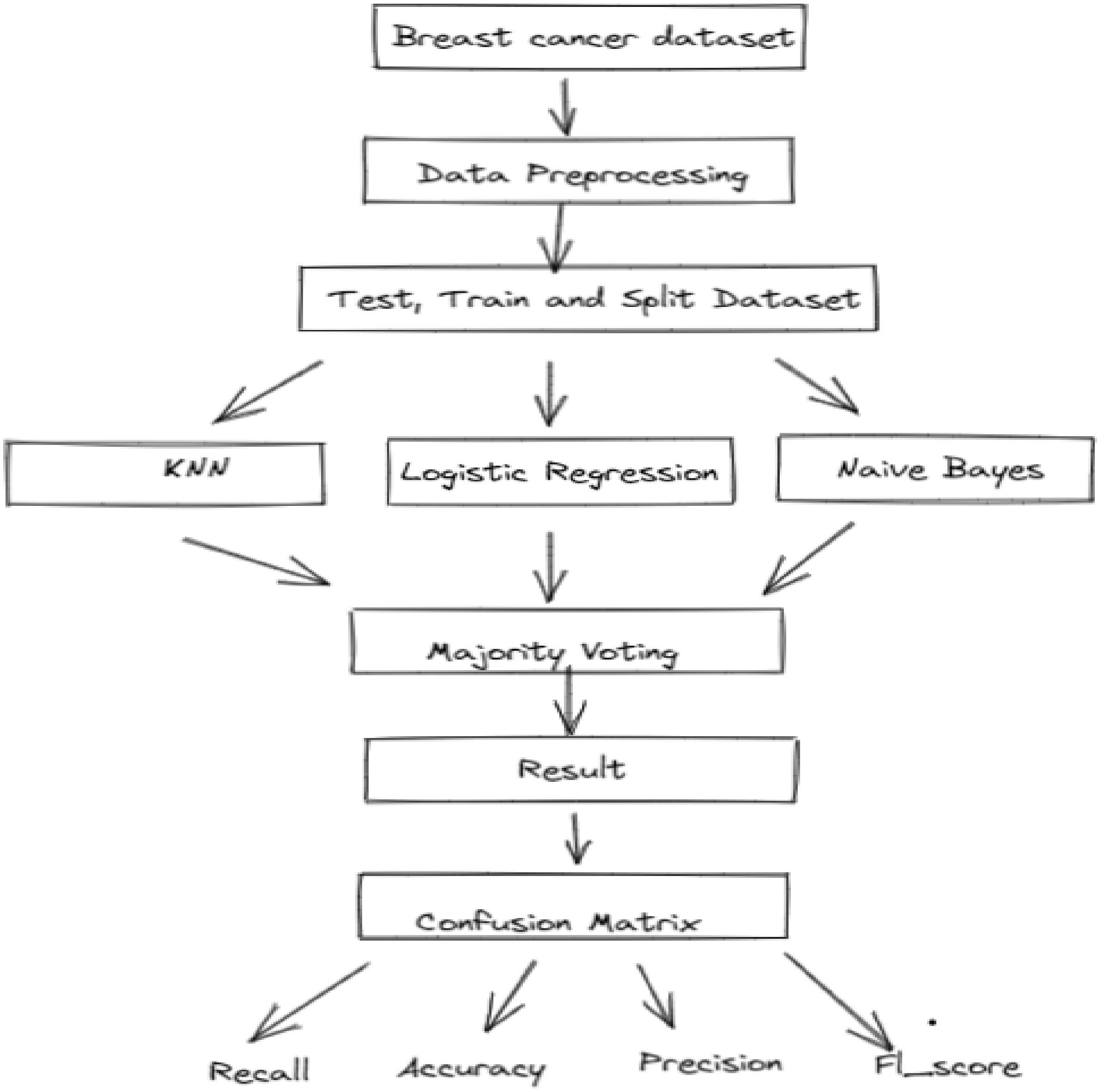


Fig. 1 Detailed Design

4.1.2 About Dataset

Breast-Cancer wisconsin Dataset imported from sk learn library

The data set has 569 data points and 30 features.

The classification needs to be done in two classes: “benign” and “malignant”.

The problem is being solved using supervised learning.

The dataset has Class Distribution: 212 - Malignant, 357 - Benign.

Malignant(cancer) is represented by zero and Benign is represented by 1.

4.1.3. Dataset preprocessing and Train-test Split:

The dataset is divided into testing and training data using the train_test_split function imported from sklearn.model_selection.

The testing data size is set to 20%, which signifies 20% data is for testing the model and 80% is for training the model.

The random state has been set which signifies that the division between the dataset will be the same every time.

4.1.4. Algorithms

4.1.4.1 Logistic Regression:-

Logistic regression is a Supervised Learning algorithm. Logistic regression is used for solving classification problems.

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not.

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.

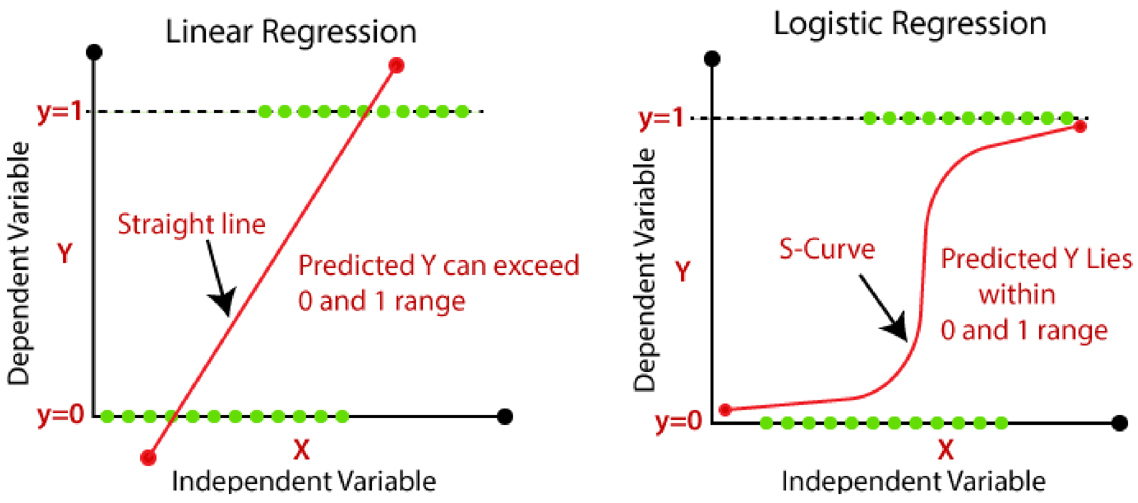


Fig 2. Logistic Regression

Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Assumptions for Logistic Regression:

- The dependent variable must be categorical in nature.
- The independent variable should not have multicollinearity.

Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

- We know the equation of the straight line can be written as:

$$y= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{\mathbf{y}}{\mathbf{1-y}}\ \textbf{; 0 for y= 0, and infinity for y=1}$$

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$\boldsymbol{log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n}$$

The above equation is the final equation for Logistic Regression.

4.1.4.2 K-Nearest Neighbour:-

- KNN is a supervised learning algorithm.
- KNN algorithm is used to solve classification problems. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using K- NN algorithm.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- In binary classification the result is predicted in the form of 0/1.
- Here, nearest neighbors are those data points that have minimum distance in feature space from our new data point. And K is the number of such data points we consider in our implementation of the algorithm.

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset.

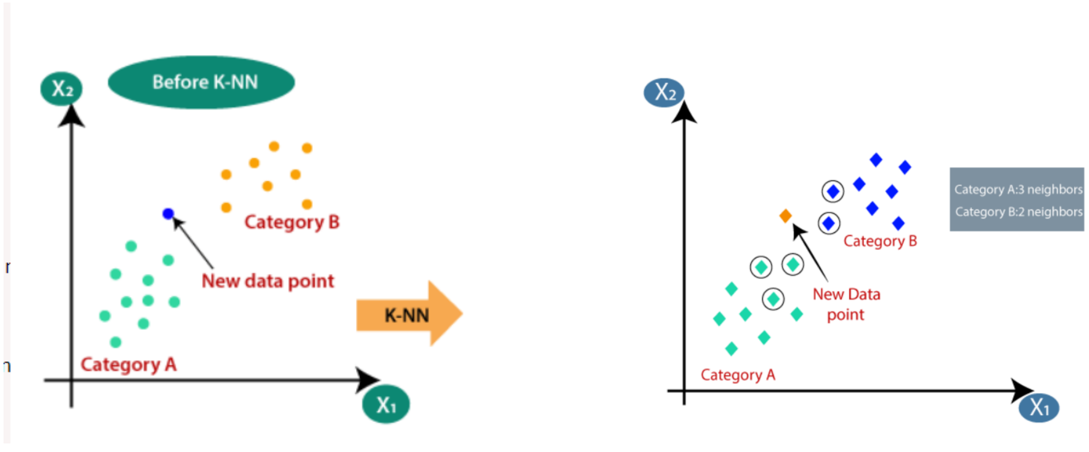


Fig 3 a. KNN

In order to predict the class of the new datapoint, we will be taking the distances of all the data points from all the training data and computing the nearest k point distances.
Then will be taking the majority voting and predicting the class of the new data point.

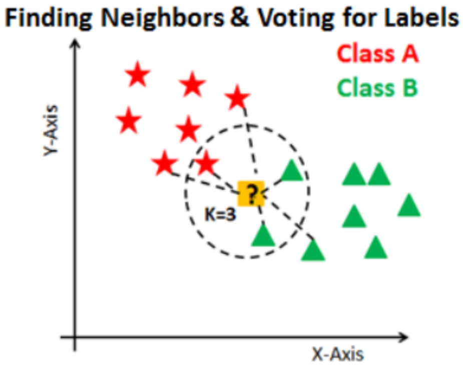


Fig 3 b. Computing K-nearest distances

4.1.4.3 Naive Bayes Classifier :-

Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature, used in a wide variety of classification tasks.

The fundamental Naïve Bayes assumption is that each feature makes an:

- Independent
- Equal

Contribution to the outcome.

What is a Classifier?

Classification is the process of predicting the class of given data points. Classes are sometimes called targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

For instance our problem statement of Breast Cancer Prediction can be classified as a classification problem. This is binary classification since there are only 2 classes "Malignant" and "Benign". A classifier utilizes some training data to understand how given input variables relate to the class. In this case, known malignant denoted by a "0" and benign denoted by a "1" emails have to be used as training data. When the classifier is trained accurately, it can be used to detect breast cancer.

Classification belongs to the category of supervised learning where the targets are also provided with the input data. There are many applications in classification in many domains such as in credit approval, medical diagnosis, target marketing etc.

Bayes Theorem:- Bayes' Theorem is a simple mathematical formula used for calculating conditional probabilities.

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Here the variable 'X' represents the features of the data set as perimeter, area radius etc. and 'y' represents the classes of data i.e. Benign and malignant.

As there are multiple features X = (x1, x2, x3, x4 …)
The equation can be rewritten as:

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

Now, you can obtain the values for each by looking at the dataset and substitute them into the equation. For all entries in the dataset, the denominator does not change, it remains static. Therefore, the denominator can be removed and proportionality can be injected.

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

In our case, the class variable(**y**) has only two outcomes, yes or no. There could be cases where the classification could be multivariate. Therefore, we have to find the class variable(y) with maximum probability.

And hence the final equation turns out to be:

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

Types of Naive Bayes Classifiers:

1. Gaussian Naïve Bayes Classifier
2. Multinomial Naïve Bayes Classifier
3. Bernoulli Naïve Bayes Classifier\

In our implementation of Breast Cancer Prediction using Machine Learning

Algorithm we have used is the **Gaussian Naive Bayes Classifier**.

Gaussian Naive Bayes Classifier:

In Gaussian Naïve Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution (Normal distribution). When plotted, it gives a bell-shaped curve which is symmetric about the mean of the feature values as shown below:

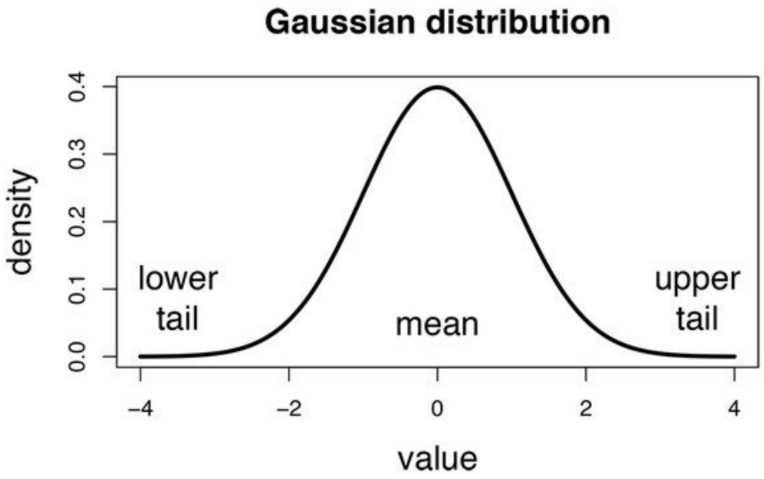


Fig 4. Gaussian Naive Bayes

The likelihood of the features is assumed to be Gaussian, hence, conditional probability is given by:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where the parameter such as Mean and Standard Deviation are represented as:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Mean

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{0.5}$$

Standard deviation

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normal distribution

The Zero-Frequency Problem

One of the disadvantages of Naïve-Bayes is that if you have no occurrences of a class label and a certain attribute value together then the frequency-based probability estimate will be zero. And this will get a zero when all the probabilities are multiplied.

An approach to overcome this ‘zero-frequency problem’ in a Bayesian environment is to add one to the count for every attribute value-class combination when an attribute value doesn’t occur with every class value.

4.1.5.Majority Voting:-

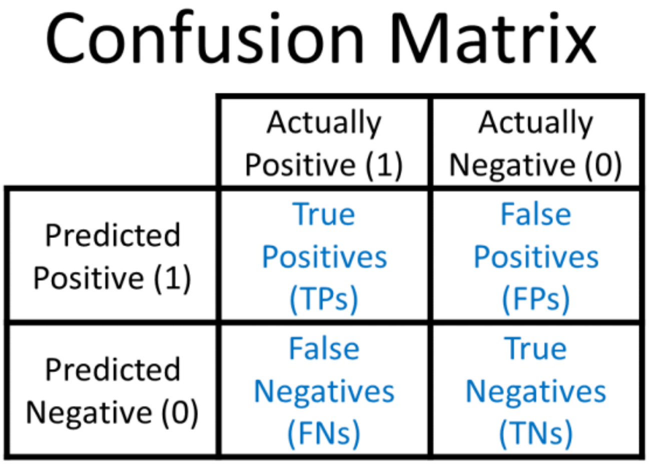
- Created 2 counter variables: counter_zero and counter_one which denotes malignant and benign respectively.
- Majority voting by taking the most voted class: malignant or benign, based on the result received by applying individual algo on the dataset.
- Storing the most voted result for each data point
- Calculating the final accuracy by using these predicted values.

4.2.Evaluation matrix

4.2.1 Confusion matrix:-

It is a table that is used in classification problems to assess where errors in the model were made.It is a table with 4 different combinations of predicted and actual values.

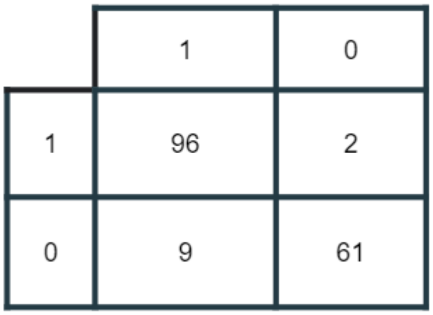
It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC-ROC curves.



Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Fig 5. Confusion Matrix



	1	0
1	96	2
0	9	61

-Fig 6. Confusion Matrix of Proposed Model

4.2.2 Accuracy :-

From all the classes (positive and negative), how many of them we have predicted correctly.Accuracy should be as high as possible.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

4.2.3 F1-Score:-

F-score helps to measure Recall and Precision at the same time.

$$\text{F1-Score} = 2 * \frac{(\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$$

4.2.4 Recall:-

Recall is from all the positive classes, how many we predicted correctly.Recall should be as high as possible.

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{\text{Predictions Actually Positive}}{\text{Total Actual positive}}$$

4.2.5 Precision:-

Precision is from all the classes we have predicted as positive, how many are actually positive.Precision should be high as possible.

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{\text{Predictions Actually Positive}}{\text{Total Predicted positive}}$$

CHAPTER - 5

IMPLEMENTATION AND RESEARCH

This proposed method for breast cancer detection consists of four main parts:

- Data Preprocessing techniques
- Data visualization
- Machine learning algorithms and
- Final Accuracy and Confusion Matrix

In this work data we extracted 569 data points and 30 features for patients who have breast cancer and the testing stage was applied on many people either they have breast cancer or they have not. In this work, a supervised learning algorithm has been used. Indeed, we used three types of supervised machine learning algorithms which were the Naive Bayes Classifier, Logistic Regression and KNN and results were extracted from all of them.

At first the Data is preprocessed to remove the null values and irrelevant features. After that data visualization is done by plotting the correlation matrix between all the features from which features with maximum correlation are taken to make scatter and pairplot graphs. This has been done using Seaborn and Matplotlib libraries.

Following this we have implemented each machine learning algorithm i.e. Naive Bayes Classifier, Linear Regression and K-Nearest Neighbour from scratch to predict if the cancer according to the features in a person is malignant or benign denoted by “0” and “1” respectively. For implementation of which we have used numpy, pandas, and collection libraries have been used. Prediction array from each algorithm is traversed and a number of malignant and benign results have been counted. This count is used to find out the mean which is our final test score. Followed by Confusion matrix along with calculation of accuracy, precision, recall and f_score are calculated and printed.

5.1 TABLE OF MODELS

Algorithms	Accuracy	Precision	Recall	F_score
KNN	0.96	0.95	0.96	0.97
NAIVE BAYES	0.97	0.94	0.92	0.96
LINEAR REGRESSION	0.95	0.95	0.94	0.95

5.2 RESULT AFTER MAJORITY VOTING

A. **Result from each algorithm**
Naive Bayes Classifier: 0.92
K-Nearest Neighbour: 0.92 (3 neighbors)
Linear Regression: 0.91

B. **Result after majority voting**
Test_Score : 0.94

C. **Confusion Matrix**
Test_Score : 0.94

5.3 PREDICTED

Accuracy	Precision	Recall	F_score
0.94	0.94	0.96	0.95

CONCLUSION

This project is implemented using machine learning techniques that are helpful in predicting Breast cancer type and assists oncologists in decision making for breast cancer patients.

For this purpose we have implemented

1. KNN.
2. Logistic Regression.
3. Naïve Bayes algorithm.

using Python in Google Colaboratory.

The experimental results show that our model performs better and provides better accuracy in predicting the breast cancer type as benign and malignant.

FUTURE SCOPE

The project can further be developed into a web or app interface which can be used by any user to input certain requested health details from which features can be derived by the system.

The system can then predict whether a person possesses Breast Cancer or not on the basis of the proposed system.

The accuracy of the system can be further increased by using other supervised learning algorithms such as Support vector Machine and Random forest Classifier.

By including the results of all 5 algorithms the best result can be predicted using majority voting and ensemble learning.

As of now, the accuracy of the system is 92-93% which is more than Logistic regression but less than K-nearest neighbor and Naive Bayes.

We can further work upon the other algorithms as well to increase the accuracy of the proposed model.

REFERENCES:-

1. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
2. https://en.wikipedia.org/wiki/Naive_Bayes_classifier
3. https://en.wikipedia.org/wiki/Logistic_regression
4. https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html
5. Logistic Regression From Introductory to Advanced Concepts and Applications
6. Introduction to machine learning with python o'reilly.
7. https://www.researchgate.net/publication/335984408_Breast_cancer
8. https://www.ieee.org/publications/