International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

An Ensemble Machine Learning Model for Breast Cancer

1st C.A. Dishanth
dept. of CSE

20191CSE0100

Presidency UniversityBengaluru, Karnataka

201910101302@presidencyuniversity.in

2nd Kuppala Sai Rishi

dept. of CSE

20191CSE0279

Presidency UniversityBengaluru, Karnataka

201910101281@presidencyuniversity.in

3rd Shaik Abdul Niyaz
dept. of CSE

20191CSE9012

Presidency UniversityBengaluru, Karnataka

201910100655@presidencyuniversity.in

4th C.H. Jagadeesh
dept. of CSE

20191CSE0099

Presidency UniversityBengaluru, Karnataka

201910101025@presidencyuniversity.in

5th P. Harsha Vardhan

dept. of CSE

20191CSE0414

Presidency UniversityBengaluru, Karnataka

2019101010692@presidencyuniversity.in

Dr. Ragaventhiran M.E, PH. D

Associate Professor, Dept of CSE

Presidency University

Bengaluru

Abstract:

Almost 1 crore people died from cancer last year, with breast cancer accounting for 22.6% of all cancer-related deaths worldwide (BC). In India, 14.7% of cancer cases are of the BC variety, which affects women more frequently than other cancers. Many studies have been done on the subject of early BC detection, which can aid with timely treatment initiation and a reduction in mortality. Roughly 86% of cases that are diagnosed are appropriately diagnosed. Machine learning techniques are useful for classifying data. Particularly in the realm of medicine, where those techniques are frequently utilized in diagnosis and analysis for decision-making.

This study conducted data visualization and performance evaluations of the Support Vector Classifier (SVC), Decision Tree Classifier, Gaussian Naive Bayes (GNB), K Nearest Classifier (k-NN), Logistic Regression and Linear discriminant analysis. The major goal is to assess each algorithm's efficiency and effectiveness in terms of accuracy and precision in the classification of data. Our goal is to evaluate several Machine Learning techniques for early, effective, and accurate detection.

Keywords:

Gaussian Naïve Bayes, Decision Tree, Linear Discriminate Analysis, Logistic Regression, Support Vector Machine Nearest Neighbor, XG Boost

Introduction:

Breast Cancer is a huge global health concern, and early identification is crucial. key to improving patient outcomes. With the potential for precise, effective, and economical diagnosis, machine learning algorithms have appeared as promising techniques for breast cancer detection.

Breast cancer is a critical worldwide health issue, and improving patient outcomes requires early detection. Because they can deliver precise, effective, and economical diagnostics, machine learning algorithms have emerged as feasible options for breast cancer screening.

Using datasets of pertinent patient data, machine learning algorithms can be trained to identify patterns and traits that are suggestive of breast cancer. The analysis and prediction of the presence of cancer using these algorithms is thus possible. One of the most prevalent cancers to impact women globally is breast cancer, and early detection is essential for effective treatment and higher survival rates. Because of its capacity to process vast volumes of data and extract valuable information for categorization, machine learning algorithms have becoming more and more used for breast cancer detection.

In this review, we employ four distinct machine learning algorithms to identify breast cancer: Gaussian Naive Bayes, K-Nearest Classifier, Decision Tree, Linear Discriminant Analysis.

Since it presupposes feature independence, the probabilistic technique known as Naive Bayes is straightforward and computationally effective. K-Nearest Neighbors is a non- parametric technique that categories new instances based on the distance between data points.

LITERATURE SURVEY:

Breast Cancer Detection Using Machine Learning Techniques, the author a model integrated with multiple machine learning (ML) methods, including the Support Vector Machine, Artificial Neural Network, and K-Nearest Neighbor, is proposed in this research. for an effective and accurate breast cancer diagnosis. The

Breast Cancer Classification Using Machine Learning Techniques The author of this study suggested reviewing previous research to categorize these malignancies. Medical image classification uses machine learning techniques comprised of Support Vector Classifier, KNearest Classifier Random Forest (RF). The outcomes demonstrated that the SVM attained great accuracy-roughly 94% (2021)

Random Forest to forecast breast cancer. In this study, the author suggested using a random forest to predict breast cancer. Random forest is one of several classification techniques and is an algorithm for categorizing enormous amounts of data. Random forest classification is used on breast cancer data to enhance classification performance and accuracy. Utilizing the random forest approach, the author has categorized breast cancer. The outcome in this study is more than 95% (2019)

BREAST CANCER CLASSIFICATION USING K-NEARESTNEIGHBORS In this study, the author classified the k Nearest Neighbor's method is used categorize the disease of breast cancer. Additionally, k-NN was applied for various k values, and the results' classification accuracy comparisons were made. The k-NN's realized classification accuracy is roughly 95%. In addition, the study's findings demonstrate that the author has identified that k-NN is a useful classifier for categorizing breast cancer disease. (2018)

Utilizing machine learning, classify breast cancer A Naive Bayes (NB) classifier and a K closest neighbor (KNN) classifier for the classification of breast cancer were presented by the author in this study. With the use of cross validation, the author suggested comparing the two new solutions' accuracy. Results indicate that KNN has a lower error rate than NB and provides the maximum accuracy (95.51%).

PROPOSED SYSTEM AND ADVANTAGES:

We Proposed method: Using six different machine learning algorithms to predict/detect breast cancer: Decision Tree Classification, Nave Bayes, K-Nearest Neighbor's, Support Vector Machines, and Logistic Regression, and Linear Discriminant Analysis. The objective of the method is to determine which algorithm has the best performance in detecting breast cancer. Breast cancer detection is a critical task in medical diagnosis as early detection can save lives. The six algorithms selected for this task are commonly used in medical diagnosis and have been shown to have good performance in detecting cancer. By comparing the performance of these six algorithms, we can determine which algorithm has the best performance in detecting breast cancer and can be used as a tool for medical diagnosis.

Advantages

- High accuracy.
- Time Saving.
- Low Complexities
- Easy To Scale

METHODOLOGY:

Data Collection:

We were able to gather an enormous amount of information on breast cancer conditions as well as are now employing the information in our research on the categorization of illness. Machine learning strategies are utilised in the swift and prompt detection of breast cancer function illness along with various illness because they now hold an important place in health care and aid us in recognising and categorising illness.

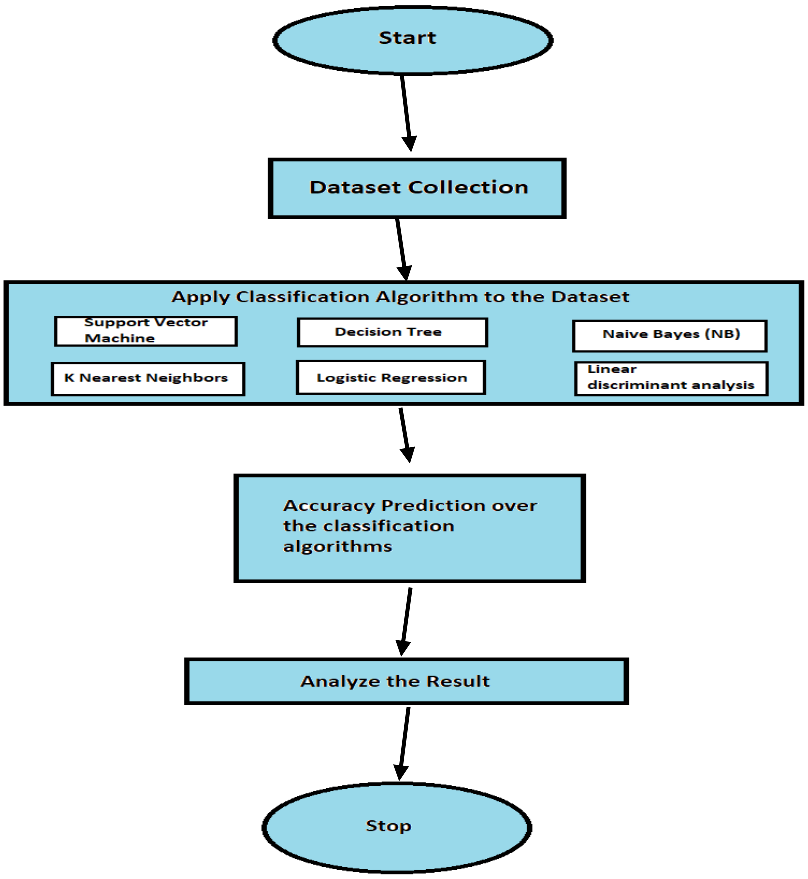
Information from outside healthcare facilities and labs that specialise in analysing and treating illnesses were collected and utilised by me for our investigation. As the information collected comprise 32 factors or qualities, all of them were used in our research to figure out

([Radius mean, texture mean, perimeter_mean,area_mean,smothness_mean,compactness_mean,concavity_mean,concave points mean, symmetry mean, fractal_dimension_mean,raduis_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_points_se, concave_points_se, symmetry_se, fractal_dimension_se, raduis_worst, texture_worst, perimeter_worst,area_worst, smoothness_worst, compactness_worst, concavity_worst, concavepoints_worst, Symmetry_worst,fractual_dimension_worst)]

Data Pre-processing:

Pre-processing the information is a crucial stage in data analysis because it makes a positive impact on the information. The pre-processing method is applied to disclose the material by analysing it as well as finding what was deleted because it carefully investigates the information. Data preparation and cleansing are all part of the pre-processing phase. We cleaned & organised the information that they had the opportunity to get in that phase or gradually and we also discovered a set of missing information where the absent characteristics are. We were capable of to analyse the data that was missing through. Additionally, we combined the MLP algorithm with normalisation techniques.

ARCHITECTURE DIAGRAM:



Algorithms:

- Gaussian Naïve Bayes
- Decision Tree
- Linear Discriminant Analysis
- K Nearest Neighbor
- Support Vector Machine
- Logistic Regression
- XG Boost

Gaussian Naive Bayes:

Gaussian Naive Bayes (Gaussian NB) is a variant of the Naive Bayes algorithm that assumes that the features (input variables) follow a Gaussian distribution. Like other Naive Bayes algorithms, it is a probabilistic algorithm used for classification tasks. GaussianNB can be trained on labeled data to become familiar with each feature's mean and standard deviation for each class. Then, given a new input, it can calculate the probability of each class given the observed feature values, using the Gaussian probability density function.

Advantages of Gaussian NB:

GNB is a basic and easy-to- understand that algorithm is relatively easy to implement. It is computationally efficient and can handle large datasets with high dimensionality.

Formula of GNB:

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

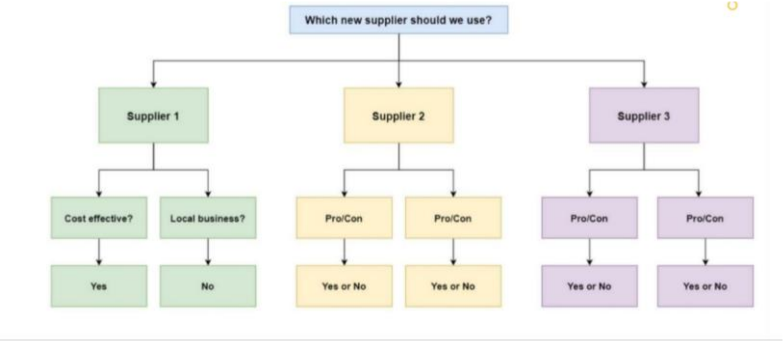
Decision Tree:

Each leaf node of a decision tree that resembles a flowchart signifies the outcome, while an internal node stands in for an attribute and a branch for decision rule. Decision trees have a root node positioned at the tree's top. By considering attribute values, data subsets can be formed. Recursive partitioning divides the tree into distinct segments, resembling a flowchart. This flowchart-like structure aids in decision-making, mirroring the thought process of individuals may be easily understood and interpreted as a result of this.

Advantages of decision tree classifier:

Decision trees provide a clear and simple way to understand the decision-making process. Decision trees can handle large amounts of data quickly and efficiently, making them suitable for analyzing medical data with many features. Decision trees can be updated easily as new data becomes available, allowing for the improvement of the predictive accuracy over time.

Working of decision tree:



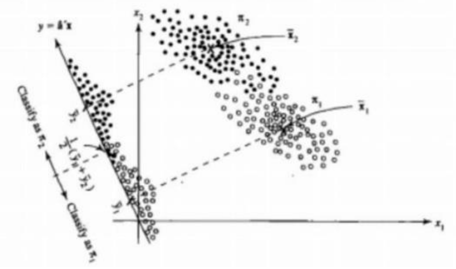
Linear Discriminant analysis:

Data are divided into distinct categories or classes using the statistical approach known as linear discriminant analysis (LDA). It is a method of dimensionality reduction that pinpoints the key characteristics in the data that distinguish between classes. By combining the data in a linear way, LDA is able to maximize the distance between classes while minimizing the variation inside each one. It presumes that both the covariance matrices for each class are equal and that the data is normally distributed. LDA is frequently used in classification tasks in machine learning. pattern recognition, and image processing applications

Advantages of using linear discriminant analysis:

LDA is an efficient algorithm, making it suitable for analyzing large medical datasets. It makes efficient use of training data by modeling the distribution of each class, which allows for better classification accuracy with fewer training samples. LDA can be extended to handle multi-class problems, where there are more than two classes. This can be useful in breast cancer detection

Working of LDA:



K Nearest Neighbor:

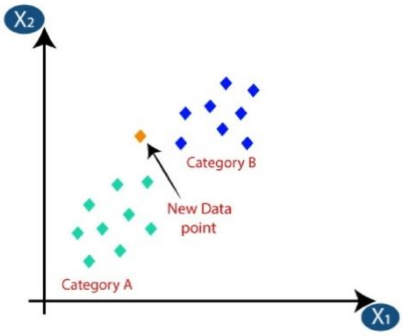
The majority class of a new data point's k nearest neighbors in the training data determines its class in the k-NN classifier. A hyperparameter called k must be established before the model is trained. Any distance metric, including

Manhattan and Euclidean, can be used to identify the closest neighbors. The k-NN approach is appropriate for both linear and non- linear interactions between the features and the target variable and makes no assumptions about the underlying distribution of the data.

Advantages of K Nearest Neighbor:

KNN does not require a training phase, allowing for quick analysis of new data. K Nearest can be used as a base classifier in ensemble methods such as bagging and boosting, improving the predictive accuracy of the model. KNN (K-Nearest Neighbors) is a versatile algorithm that can be applied to both classification and regression tasks, making it suitable for analyzing medical data.

Working of K Nearest Neighbor:



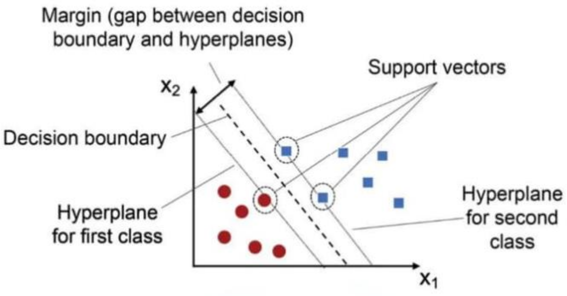
Support Vector Machine:

SVM, the goal is to find the best possible boundary (or hyperplane) that can separate different classes of data points in a high-dimensional space to divide various classes, SVM creates a hyperplane in multidimensional space. To reduce error, SVM builds an ideal hyperplane in an iterative fashion. The main goal is to identify the largest marginal hyperplane that best classifies the dataset. SVM models are essentially hyperplane representations of several classes in multidimensional space. In order to lower error, SVM will iteratively construct the hyperplane. To find a maximum marginal hyperplane, SVM divides the datasets into classes.

Advantages of Support Vector Machine:

SVM is known to have high accuracy in classification tasks, and breast cancer detection is no exception. SVM can effectively distinguish between malignant and benign tumors, resulting in accurate predictions. SVM produces interpretable results, which can help clinicians understand the features that are most important in predicting tumor classification. This can assist in developing treatment plans and monitoring patient progress. Overall, SVM is a powerful algorithm that can be used for accurate and efficient breast cancer detection.

Working of SVM:


Margin (gap between decision boundary and hyperplanes)
Support vectors
x_2
Decision boundary
Hyperplane for first class
Hyperplane for second class
x_1

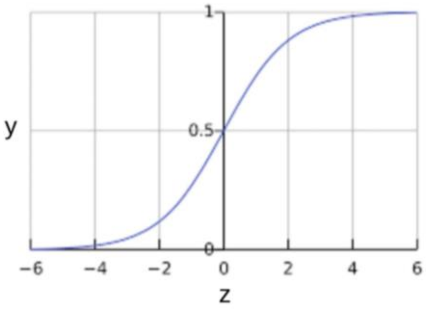
Logistic Regression:

The objective of logistic regression, a data analysis approach, is to discover relationships between two data components. It utilizes this relationship to predict the value of one parameter based on the other. The predicted outcome typically falls into a limited number of options, such as "yes" or "no."

Advantages of Logistic Regression:

It is effective when working with small datasets, which is often the case with medical datasets such as breast cancer detection. Logistic Regression produces interpretable results, which can help clinicians understand the relationship between different features and tumour classification. This can assist in developing treatment plans and monitoring patient progress. Overall, Logistic Regression is a reliable and simple algorithm that can be used for accurate breast cancer detection, especially when working with small datasets.

Working of Logistic Regression:



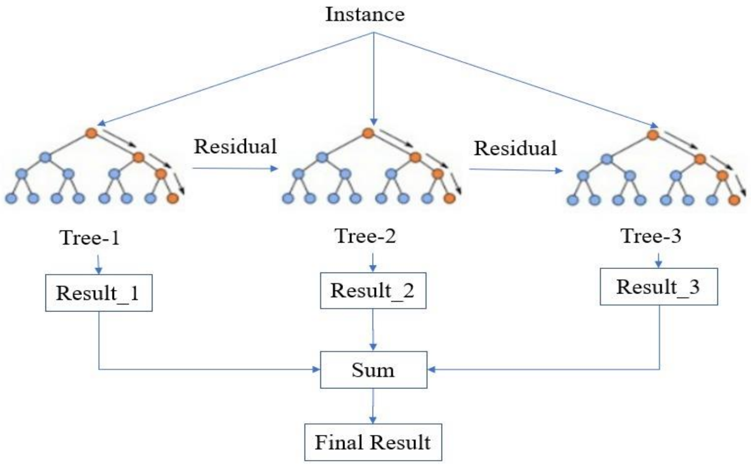
XG BOOST:

In order to produce more precise forecasts, it combines a number of straightforward decision trees. Each decision tree is constructed one after the other, with each new tree attempting to fix the mistakes produced by the prior tree. To raise the accuracy of each decision tree, XGBoost employs a gradient boosting technique. To produce more precise predictions, it can manage missing values in the data and determine the significance of each attribute. XGBoost is renowned for handling huge datasets with good accuracy and efficiency. It has gained popularity in a variety of fields, such as picture and speech recognition, financial forecasting, and medical diagnosis like the prediction of breast cancer.

Advantages of XG Boost:

In real-world datasets, missing values are common. XGBoost has a built-in feature to handle missing values, which can improve its accuracy in cases where other models may struggle. XGBoost can calculate the importance of each feature in the dataset, which can help to identify which features are most relevant for breast cancer prediction.

Working of XG Boost:



NO	Algorithms	Accuracy
1	Gaussian Naïve Bayes	94.14%
2	Decision Tree	92.26%
3	Linear Discriminate Analysis	95.78%
4	K Nearest Neighbor	96.48%
5	Support Vector Machine (SVM)	97.90%
6	Logistic Regression	98.12%
7	XG Boost	96.47%

Conclusion:

One of the illnesses that affects everyone worldwide and is becoming more prevalent is breast cancer disease. Our study examines the categorization of breast cancer disease between hyperthyroidism and hypothyroidism in light of medical reports that demonstrate substantial abnormalities in breast cancer diseases. Algorithms were used to classify this illness. Using multiple methods, machine learning produced favourable outcomes. The accuracy of the Logistic Regression produced a value of 98.12% in the model. This study emphasises the value of early identification in enhancing patient outcomes and the significance of breast cancer prediction. We have made significant progress in creating precise predictive models for breast cancer diagnosis by utilising cutting-edge machine learning algorithms and data mining techniques. These models make good use of a wide range of variables, such as demographic data, clinical traits, genetic markers, and imaging data, to identify individuals who are at risk and enable prompt interventions.

References:

[1] Breast Cancer Detection Using Machine Learning Techniques International Journal for Research in Applied Science & Engineering Technology (IJRASET)May 2022-Sarthak Vyas, Abhinav Chauhan, Deepak Rana, Noman Ansari, Meerut Institute of Engineering and Technology

(MIEThttps://doi.org/10.22214/jjraset.2022.43055)

[2] Breast Cancer Classification Using Machine Learning Techniques Turkish Journal of Computer and Mathematics Education (TURCOMAT) 2021 SeptemberSrwa Hasan University of Sulaimani

https://www.researchgate.net/publication/356844442

[3] Random Forest for breast cancer prediction T. L Octaviani and Z. Rustama Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA). University of Indonesia, (2019);

https://doi.org/10.1063/1.5132477

[4] BREAST CANCER CLASSIFICATION USING K- NEAREST NEIGHBORS ALGORITHM he Online Journal of Science and Technology - July 2018 Istanbul Commerce University, Department of Computer Engineering. Istanbul-Turkey

www.tojsat.net

[5] Breast Cancer Classification Using Machine Learning Meriem AMRANE Saliha OUKID Computer Science Department, LRDSI Laboratory, University of Blida, 2017

Algeriahttps://doi.org/15.5863/1.342477