**Nadilla Yaswanth Baba BL.EN.U4AIE20039**

**REPORT**

**BREAST CANCER PREDICTION USING MACHINE LEARNING**

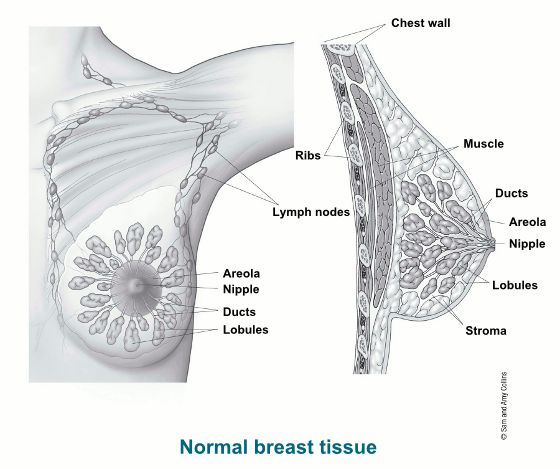
**INTRODUCTION**

* 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.

**What is Breast Cancer?**

Breast cancer is a type of cancer that starts in the breast. It can start in one or both breasts. Cancer starts when cells begin to grow out of control. Breast cancer occurs almost entirely in women, but men can get breast cancer, too.

**It’s important to understand that most breast lumps are benign and not cancer (malignant).**Non-cancer breast tumours are abnormal growths, but they do not spread outside of the breast. They are not life threatening, but some types of benign breast lumps can increase a woman's risk of getting breast cancer. Any breast lump or change needs to be checked by a health care professional to find out if it is benign or malignant (cancer) and if it might affect your future cancer risk.



* Breast cancer treatment can be highly effective, especially when the disease is identified early.
* But the diagnosis is very expensive, and the mental pressure would be immense on the patient.
* So, when a doubt arises, instead of taking the tests directly we can use these machine learning models to predict the chances that we are prone to cancer based on the authentic data sets of previously diagnosed patients. As a result, we can save a lot of money and valuable time.

**Why Machine-Learning?**

Diagram

Description automatically generated

**ML Algorithms used in our Project**

* **Logistic Regression**
* **K-Nearest Neighbors**
* **Support Vector Machine - LINEAR**
* **Support Vector Machine - RBF**
* **NAÏVE BAYES**
* **DECISION TREE**
* **RANDOM FOREST**

Now, let us look at each Algorithm in detail…

1. **Logistic Regression**

* It is a statistical method for analyzing the dataset with one or more independent variables
* It is used to predict a binary outcome (1/0), (yes/no), (true/false) etc. given a set of independent variables
* But not limited to binary outcome alone, it can also be used to predict whether the unseen data belongs to Class A or not Class A for multi class problems
* Though it uses the term regression, it is applied for classification problems only
* For binary outcome
  + Whether fraud/not, spam/not
* For non-binary outcomes
  + Whether in class Red or in class Green.

[classes Green/Blue]

Diagram

Description automatically generated

* Logistic function / sigmoid function is represented as
* where
* , .. are the regression coefficients and
* where P denotes the maximum likelihood of an occurrence of an event

1. **K-Nearest Neighbors**

* K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most like the available categories.
* 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 suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

Chart, scatter chart

Description automatically generated

The K-NN working can be explained on the basis of the below algorithm:

* Step-1: Select the number K of the neighbors
* Step-2: Calculate the Euclidean distance of K number of neighbors
* Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
* Step-4: Among these k neighbors, count the number of the data points in each category.
* Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

1. **Support Vector Machine - LINEAR.**

* Support vector machines are supervised machine learning algorithms used for regression and classification analysis.
* The objective of the algorithm is to fit the best hyperplane in an N-dimensional space that classifies data points into separate classes.
* Support vectors are the data points that are close to hyper plane and can influence its position and orientation.

Chart, scatter chart

Description automatically generated

* Linear SVM can be used when we have data that is linearly separable
* Example , a 2-D data that has 2 classes
* Here the SVM fits best straight line that maximizes margin between support vectors.

Chart, scatter chart

Description automatically generated 

1. **Support Vector Machine – Radial Basis Function**

* The RBF kernel region of SVM is also actually a linear decision-making boundary as well
* But what it does is it uplifts your linearly non separable data into higher dimension and try to project a linear boundary to separate your classes.

A picture containing text, watch

Description automatically generatedChart, scatter chart

Description automatically generated

1. **Naïve bayes**

* Naive Bayes is one of the most popular supervised learning classifications that is used for the analysis of the categorical text data.
* It is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.
* Naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is, each feature being classified is not related to any other feature. The presence or absence of a feature does not affect the presence or absence of the other feature.
* It is important to understand the Bayes theorem concept first as it is based on the latter.
* Bayes theorem calculates the probability of an event occurring based on the prior knowledge of conditions related to an event. It is based on the following formula:
  + - **P(A|B) = P(A) \* P(B|A)/P(B)**
  + Where we are calculating the probability of class A when predictor B is already provided.
  + P(B) = prior probability of B.
  + P(A) = prior probability of class A
  + P(B|A) = occurrence of predictor B given class A probability

1. **Decision Tree**

* Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.
* Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.
* We can represent any Boolean function on discrete attributes using the decision tree
* At the beginning, we consider the whole training set as the root.
* Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
* Based on attribute values records are distributed recursively.
* We use statistical methods for ordering attributes as root or the internal node.

1. **Random Forest**

* As the name suggests, Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset*.*
* Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
* The greater number of trees in the forest leads to higher accuracy.
* Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.
* The Working process can be explained in the below steps:
  + Step-1: Select random K data points from the training set.
  + Step-2: Build the decision trees associated with the selected data points (Subsets).
  + Step-3: Choose the number N for decision trees that you want to build.
  + Step-4: Repeat Step 1 & 2.

**Results**

Text, letter

Description automatically generated

Calendar

Description automatically generatedCalendar

Description automatically generated

Table

Description automatically generated

**THANK YOU!**