**ARTICLE**

**ANALYSIS OF BREAST CANCER PREDICTION USING MACHINE LEARNING**

**ANANYA PAGADALA,2203A52046**

\***Affiliation1:2203a52046@sru.edu.in**

\* **SR UNIVERSITY**

\* **Section:AIML-C**

**ABSTRACT:** This study explores different regression models to forecast diseases, diving into the field of predictive analytics. The goal of the study is to compare the effectiveness of four distinct approaches: K-Nearest Neighbours (KNN), Preceptron learning, Support Vector Machine (SVM), and Logistic Regression. Finding the best model to forecast diseases based on characteristics such as age, gender, proteins, cancer stage, histology, ER status, PR status, HER2 status, kind of surgery, and patient status is the main goal. Combining thorough evaluation, model training, and data preparation, the study offers insightful information about the advantages and disadvantages of each method.

**KEYWORDS:** Logistic regression; Support Vector Machine; Perceptron learning ; K-Nearest Neighbors.

**I .INTRODUCTION:**

Breast cancer is a common, possibly fatal illness that predominantly affects women, however it can strike men as well. It is typified by the unchecked proliferation of cancerous cells in the breast tissue, which may result in tumour development. Age, genetics, and lifestyle are some of the different risk factors that contribute to breast cancer, which is the most frequent cancer among women globally. Improving the prognosis and results of treatment depends on early detection, which can be achieved by routine breast self-examinations, mammography, and physical examinations. Depending on the unique features of the malignancy, treatment options could include radiation therapy, chemotherapy, surgery, and targeted therapies.

My data set contains 334 rows and 15 columns.

The data set consists of 15 dimensional variables:

1. Age
2. Gender
3. Protein 1
4. Protein 2
5. Protein 3
6. Protein 4
7. Tumour stage
8. Histology
9. ER status
10. PR status
11. HER2 status
12. Surgery type
13. Date of surgery
14. Date of last visit
15. Patient status

**II . Literature Review**

Research on breast cancer prediction is essential if we are to improve patient outcomes and early detection. To create prediction models, a variety of data-driven and machine learning techniques have been used. Typically, these models use clinical history, genetic markers, mammography results, and patient demographics to determine a person's risk of breast cancer.

In terms of sensitivity and accuracy, recent research has produced encouraging results that could lead to the development of more accurate risk assessment instruments. Furthermore, studies have looked into combining deep learning and artificial intelligence methods to improve prediction models even more. Personalised screening and treatment choices may benefit from these methods.

Research on breast cancer prediction has several obstacles, such as the requirement for extensive and varied datasets, handling ethical and biassed concerns, and guaranteeing the practicality of predictive algorithms. The field's continued development could lead to improved early detection and preventive methods for breast cancer, which would lessen the disease's overall impact on people and health care systems.

**METHODOLOGY:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Gender | Protein1 | Protein2 | Protein3 | Protein4 | Tumour stage | Histology | ER  status | PR  status |
| 42 | Female | 0.95256 | 2.15 | 0.0079716 | -0.04834 | II | Infiltrating Ductal Carcinoma | Positive | Positive |
| 54 | Female | 0 | 1.3802 | -0.49803 | -0.50732 | II | Infiltrating Ductal Carcinoma | Positive | Positive |
| 63 | Female | -0.52303 | 1.764 | -0.37019 | 0.010815 | II | Infiltrating Ductal Carcinoma | Positive | Positive |
| 78 | Female | -0.87618 | 0.12943 | -0.37038 | 0.13219 | I | Infiltrating Ductal Carcinoma | Positive | Positive |

Age- (Role- Target, Type- Categorical, Missing Values- No)

Gender- (Role- Feature, Type- Continuous, Missing Values- No)

Protein1- (Role- Feature, Type- Continuous, Missing Values- No)

Protein2 - (Role- Feature, Type- Continuous, Missing Values- No)

Protein3 - (Role- Feature, Type- Continuous, Missing Values- No)

Protein4- (Role- Feature, Type- Integer, Missing Values- No)

Tumour stage - (Role- Target, Type- Continuous, Missing Values- No)

Histology - (Role- Feature, Type- Continuous, Missing Values- No)

ER status - (Role- Feature, Type- Continuous, Missing Values- No)

PR status - (Role- Feature Type- Continuous, Missing Values- No)

HER2 status - (Role-Feature,Type-Continuous, Missing Values- No)

Surgery type - (Role-Feature, Type- Continuous, Missing Values- No)

-Output variable:

Class- categorical classified into 3 –[1,2,3]

**MY IMPLIMENTATIONS IN MY DATA SET:**

1. **LOGISTIC REGRESSION:**

It is a binary classification, if the target values are categorical then we can apply supervised learning. We apply logistic regression.

Logistic regression follows the parametric method approach.

Following equation with known and unknown variables.

Y=mx+c

Z=w0+w1x1+w2x2…….wnxn

Z values are continuous values so, to convert catogirical we require categorical function.

-ylog(yp)-(1-y)log(1-yp)

Where,

(Variance) Yp=1/1+e^-z =>sigmoid function

In logistic regression the activation function is sigmoid

The output sigmoid equation contains only 0 to 1

In logistic regression we will calculate cress entropy loss i.e log error.

**2 PERCEPTRON:**

One component of an artificial neural network is the perceptron. A linear machine learning approach called a perceptron is used to learn different binary classifiers under supervision. Neural networks can learn new elements and process them one at a time while preparing thanks to this technique.

**3.SUPPORT VECTOR MACHINE(SVM):**

Support Vector Machine” (SVM) is a supervised machine learning algorithm. Which can be used for both classification or regression challenges. How ever, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have)we perform classification by finding the hyper-plane that differentiates the two classes very well.

**4.K-NEAREST NEIGHBOUR(KNN):**

k-Nearest Neighbour is a non-parametric method approach finds k -nearest neighbours from the all data points. In this method our target is to find optional k-value for the given data set. K value may be k=1,2,3—n here n is no of rows. In KNN we will calculate distances between given data points to all data points from the given data set. after calculating distances we will select KNN and the average y-value of KNN is the y-value of given data point. For calculating the distances we can use Eucledian or manhatlern distance |x2-x1|+|y2-y1|

1. **RESULT:**

Firstly, I selected my data set and I did reading of my data set,the values which I don’t want to dropped them and did the implementations related to data set.I have implemented logistic regression,support vector machine(SVM),K-nearest neighbour(KNN),perceptron learning.

|  |  |  |
| --- | --- | --- |
| S.NO | Model | Accuracy |
| 1 | Logistic Regression | 1.0 |
| 2 | Support Vector Machine | 1.0 |
| 4 | KNN | 0.8659217877094971 |
| 3 | Perceptron | 1111111111111…. |

**CONCLUSION:**

Our analysis of the wine data set unveiled compelling insights. Our journey underscores the power of the data\_driven decisions through machine learning, showing the simplicity ant efficiency of the machine learning models on determining the “BREAST CANCER”.

**ACKNOWLEDGEMENT:**

My special thanks to D.RAMESH sir on guiding and supporting me.

**REFERENCE:**

**\***UCI Machine Learning Respiratory(Google).

**\***Kaggle website(Google)