A report on
Breast cancer prediction

Submitted by: Ichchha Babu Bhattarai

Date of submission June 7,2023

TABLE OF CONTENT

Abstract

Breast cancer is a prevalent and life-threatening disease affecting millions of women worldwide. Early detection and accurate prediction of breast cancer are vital for effective treatment and improved patient outcomes. This report presents a comprehensive study on breast cancer prediction using machine learning algorithms like logistic regression and support vector machines . Leveraging a well-known machine learning dataset comprising 569 rows and 31 columns, we develop and evaluate predictive models. Among the 31 columns, 30 contain the feature values and the last column is the target value containing the label which contains the encoded value of 0 and 1 . 0 represents the Malignant, and 1 represents the Benign one. In the context of breast cancer, the term "malignant" refers to cancerous tumors, while "benign" refers to non-cancerous tumors.

The results obtained from both classifiers are observed and it was found that SVM gives a bit better accuracy than the logistic regression . But a model built using both classifiers can be taken as a good model as both models give accuracy over 90% in both training and testing dataset and there is no overfitting problem too.

Introduction

A neoplasm is an abnormal mass of tissue, the growth of which exceeds and is uncoordinated with that of the normal tissues, and persists in the same excessive manner after cessation of the stimulus which evoked the change. Cancer can start almost anywhere in the human body, which is made up of 37.200 billion cells. As these tumors grow, some cancer cells can break off and travel to distant places in the body through the blood or the lymph system and form new tumors far from the original one. Unlike malignant tumors, benign tumors do not spread into, or invade, nearby tissues. Breast cancer refers to a pathology in which a tumor develops in the breast tissue. Breast cancer is amongst the most common type of cancer in both sexes since 1975 and causes around 411,000 annual deaths worldwide. It is predicted that the incidence for worldwide cancer will continue to increase, with 23,6 million new cancer cases each year by 2030, corresponding to 68% more cases in comparison to 2012. The utilization of data science and machine learning approaches in medical fields proves to be prolific as such approaches may be considered of great assistance in the decision making process of medical practitioners. With an unfortunate increasing trend of breast cancer cases, comes also a big deal of data which is of significant use in furthering clinical and medical research, and much more to the application of data science and machine learning in the aforementioned domain.

Objective

The main objective of this project is
to predict whether a breast cancer cell is Benign or Malignant

Literature review

Wang (2016) applied the ReliefF algorithm to select relevant genes associated with breast cancer. They found that a reduced set of genes achieved comparable prediction accuracy to the full gene set. Other studies have investigated the integration of clinical, imaging, and genomic data to enhance feature representation and improve prediction performance.

Al-Dulaimi . (2019) utilized an SVM model with a radial basis function kernel to predict breast cancer recurrence. Their results showed high accuracy and improved prediction compared to traditional statistical methods.

Ramírez-Gallego et al. (2017) proposed a hybrid ensemble model combining bagging and boosting with random forest and extreme learning machines. The hybrid ensemble achieved superior prediction performance compared to individual models.

Methodology

1 Data collection

The report covers the Breast Cancer Wisconsin (Diagnostic) DataSet) in kaggle created by Dr. William H. Wolberg, a physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. The data used for this project was collected in 1993 by the University of Wisconsin and is composed of the biopsy result of 569 patients in Wisconsin Hospital. Using the sklearn library the dataset is loaded in the working environment.

2. Data preprocessing

The data is checked whether it has missing value or not .It is the standard dataset so there are no missing values.

3. Features

Attribute Information: 1. ID number 2. Diagnosis (M = malignant, B = benign) •

Ten features were computed. They are
1. radius
2. Texture
3. perimeter
4. area
5. smoothness
6. Compactness
7. Concavity
8. concave points
9. symmetry
10. fractal dimension

The mean, standard error and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 variables.

4 Model selection

Since it is a classification problem model supporting it are used.Both logistic regression approach and support vector machine approach is used.

5 Model training

After selecting the ML model, the next step is to train the model using the prepared data. The training data is divided into two sets: a training set and a testing set. The training set is used to train the model, while the testing set is used to evaluate the performance of the model.The ratio of the training and testing set is 8:2 with random shuffling.

6 Model evaluation

One of the performance metrics i.e accuracy is calculated for both training and testing set.

Model/accuracy	Train	Test
Logistic regression	94.72	96.49
SVM	97.58	92.98

The training accuracy for logistic regression is 94.72 and for SVM is 97.58 and testing accuracy is 97.58 and 92.98.

Result and conclusion

The application of machine learning algorithms in breast cancer prediction holds great promise. It has the potential to revolutionize clinical practice by providing valuable insights for healthcare professionals, enabling early detection, and facilitating personalized treatment strategies.

The model is just run on the jupyter notebook or google colab.It can be deployed in any web app and prediction can be made easily.