MINI PROJECT REPORT

Topic- Breast Cancer Prediction Using Machine Learning

(CSE IV Semester Mini Project) 2021-2022

Name : Rachit Kukreja

University Roll Number : 2015262

Class Roll Number : 58

Section : ML

Resource Person : Kireet Joshi

INDEX :

INTRODUCTION

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society.

Dataset/Source: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Structured/Unstructured data: Structured Data in CSV format

Technical Tools used : Python and Jupyter notebook

OBJECTIVES

This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this i have used machine learning classification methods to fit a function that can predict the discrete class of new input.

DATA PREPARATION

The dataset used in this story is publicly available and was created by Dr. William H. Wolberg, physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA.

Project Overview:

Data Processing:
First we import dataset into our python code. Then Using Label Encoder from Scikit Learn Library we will convert labels/Catagorical values into numbers/numeric values which our Machine Learning Algorithms better understand.

Data Visualization:
Using Matplotlib and Seaborn Libraries we visualize data and establish relationship.

Data Splitting:
Using Train_test_split . We Split the data into train and test.
In Training for the purpose of this Project I have Used 4 ML Algorithms

1. Logistic Regression:
   Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes
2. Decision Tree:
   In general, Decision tree analysis is a predictive modeling tool that can be applied across many areas. Decision trees can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions.
3. Random Forest:
   Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees
4. K-neighbors:
   In statistics, the k-nearest neighbor's algorithm (k-NN) is a nonparametric classification method first developed by Evelyn Fix and Joseph Hodges in 1951, and later expanded by Thomas Cover. It is used for classification and regression. In both cases, the input consists of the k closest training examples in data set. The output depends on whether k-NN is used for classification or regression

I Have Used These 4 ML Algorithms because these are Classification ML Algorithms and we have Classify Either Cancer is Benign or Malignant.

RESULT

Accuracy with Logistic Regression = 91.22%
Accuracy with Decision Tree = 90.35%
Accuracy with Random Forest = 97.36%
Accuracy with K-Neighbors = 93.85%

CONCLUSION

The diagnosis procedure in the medical field is very expensive as well as time-consuming. The system proposed that machine learning technique can be acted as a clinical assistant for the diagnosis of breast cancer and will be very helpful for new doctors or physicians in case of misdiagnosis. From the study, we can conclude that machine learning techniques are able to detect the disease automatically with high accuracy.