BREAST CANCER PREDICTION

Team Members:

P Subiksha(19BCE1255)

S Divyashree(19BCE1689)

Dataset:

Breast cancer wisconsin

It is a dataset of Breast cancer patients with Malignant and Benign tumor.

ABSTRACT:

Breast cancer occur in women mainly in mammary gland.This dataset about breast cancer is taken from the kaggle.This is useful for predicting many incidents in day-to-day life.Using this dataset we can predict the possibility of getting breast cancer

DESCRIPTION ABOUT THE PROJECT:

This project is about how to predict breast cancer. In the dataset we have mentioned the data to predict the possibility of getting cancer, if the person has the corresponding symptoms.

We have used logistic regression and SVM algorithm to predict the cancer and found the accuracy.

In these two algorithms, the accuracy is more for SVM.

DESCRIPTION:

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. n the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

Breast Cancer Wisconsin (Diagnostic) Data Set is used in predicting whether the cancer is benign(B) or malignant(M) using neural network and decision tree.Neural networks can be used effectively to classify samples, i.e., map input data to different classes or categories. Here, will use the neural networks to classify the samples into two classes/categories i.e M(malignant) or B(benign).

METHODOLOGY:

REGRESSION:

Regression is a statistical measurement used in finance,investing, and other disciplines that attempts to determine the strength of the relationship between one dependent variable and series of other changing variables(known as independent variables)

Regression helps investment and financial managers to value assets and understand the relationships between variables.A regression problem is when the output variable is a real or continuous value, such as “salary” or “weight”.

Eg:Predicting the age of a person is an example of regression(because it is a real value)

(i) LOGISTIC REGRESSION:

Logistic regression is basically a supervised classification algorithm. Contrary to popular belief,logistic regression is a regression model.The model builds regression model to predict the probability that a given data entry belongs to the category numbered as “1”.

CLASSIFICATION:

A classification problem is when the output variable is a category,such as “red” or “blue”.A classification model attempts to draw some conclusion from observed values.In short ,classification either predicts categorical class labels or classifies attributes and uses it in classifying new data.

Eg:Predicting the gender of a person is an example of classification.

(i) SUPPORT VECTOR MACHINE (SVM):

In machine learning,support-vector machines (SVMs,also support-vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.SVMs can efficiently perform non-linear classification using what is called the kernel trick,implicitly mapping their inputs into high-dimensional feature spaces.

ALGORITHMS FOR THE DATASET:

1.LOGISTIC REGRESSION:

Logistic regression is used to predict whether the given patient is Malignant or Benign tumor based on the attributes in the given dataset.

STEP1: Load the required libraries and Input the values for all the attributes for the given dataset (ie. loading dataset) and print them.

STEP2: In this step we are go on drop columns 'id' and unnamed:32' as they have no in prediction and initialize "M" ie malignant as '1' and "B" ie benign as '0'.

STEP3: Input some integer value to store the value of the result ie if "M" then it has to return 1 else return 0. This step is input and output data.

STEP4:Here, in this step we are going to do normalization as it is necessary and read value x to store it.

STEP 5:Next is to split the data for training and testing and get the output after splitting it.

STEP6:Initialize the value for weight and bias and define a function called sigmoid function for doing the required calculation using the formula z to check whether the outcome is 1 or 0

STEP 7:Updating parameters for weight and bias, make forward and backward propagation.

And find cost and gradients.

STEP 8:Predictions ie go on check the value for z

If z is bigger than 0.5 then our prediction is sign one and if z is smaller than 0.5 then our prediction is sign 0.

STEP 9:Define a function logistic regression function and make use of the logistic regression content to get the desired output and calculate train or test errors if any.

STEP 10:Check the results using linear_model.logisticregression and print the train accuracy and test accurary.

2.SVM ALGORITHM:

1.IMPORT all required libraries

2. IMPORT pandas and numpy for loading data and for performing data analysis operations on it.

3. IMPORT seasorn and matplotlib.pyplot for data visulisation.

4.from sklearn.decomposition IMPORT pca for feature engineering.

5.from sklearn.preprocessing IMPORT standardscaler for data scaning.

6.from sklearn.model_selection IMPORT train_test_split for splitting dataset.

7. from sklearn.svm IMPORT svc for fitting svc model.

8. IMPORT os for file operations and print all reqiured libraries loaded.

9. CHECK the files in the given input folder and load dataset into pandas dataframe.

10.CHECK the data types of all the attributes loaded into the dataframe.

11.See first few rows of the data loaded and see the last rows of the data loaded.

12.Load the predictors into dataframes and we are not choosing columns-'i-d' , 'diagonsis', 'unnamed=32'.

13.Load the target values into dataframes 'y'.

14.Convert categorical data to numerical data and use only one column for target value.

15.Call corr() on dataframe x and reduce the attributes in the x dataframe.

16.Scale the data and drop the highly correlated columns which is not useful and apply PCA on scaled data.

17. Combine PCA data and target data and set column names for the data frames and combine PCA and target data.

18.Split data for training and testing and SVM model fitting and predict values and print confusion matrix.

ACCURACY:

By plotting the graph and referring to confusion matrix, we can able to see that SVM is more accurate than logistic regression.

The accuracy of SVM for predicting breast cancer is 98.6.