BREAST CANCER DETECTION AND PREDICTION USING MACHINE LEARNING

23/11/21

ML based Major project

To effective treatment of breast cancer, we use various **machine learning algorithms** to **predict if a tumor is benign or malignant, based on the features provided by the data.**

Breast Cancer Detection and Prediction using Machine Learning

ML BASED MAJOR PROJECT

INRTODUCTION

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.

The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modeling.

SOME RISK FACTORS FOR BREAST CANCER

The following are some of the known risk factors for breast cancer. However, most cases of breast cancer cannot be linked to a specific cause. Talk to your doctor about your specific risk.

Age. The chance of getting breast cancer increases as women age. Nearly 80 percent of breast cancers are found in women over the age of 50.

Personal history of breast cancer. A woman who has had breast cancer in one breast is at an increased risk of developing cancer in her other breast.

Family history of breast cancer. A woman has a higher risk of breast cancer if her mother, sister or daughter had breast cancer, especially at a young age (before 40). Having other relatives with breast cancer may also raise the risk.

Genetic factors. Women with certain genetic mutations, including changes to the BRCA1 and BRCA2 genes, are at higher risk of developing breast cancer during their lifetime. Other gene changes may raise breast cancer risk as well.

Childbearing and menstrual history. The older a woman is when she has her first child the greater her risk of breast cancer. Also, at higher risk are:

1. Women who menstruate for the first time at an early age (before 12)
2. Women who go through menopause late (after age 55)
3. Women who've never had children

> **"You can be a victim of cancer, or a survivor of cancer. It's a mindset." — Dave Pelzer**

ROLE OF MACHINE LEARNING IN DETECTION OF BREAST CANCER

A mammogram is an x-ray picture of the breast. It can be used to check for breast cancer in women who have no signs or symptoms of the disease. It can also be used if you have a lump or other sign of breast cancer.

Screening mammography is the type of mammogram that checks you when you have no symptoms. It can help reduce the number of deaths from breast cancer among women ages 40 to 70. But it can also have drawbacks. Mammograms can sometimes find something that looks abnormal but isn't cancer. This leads to further testing and can cause you anxiety. Sometimes mammograms can miss cancer when it is there. It also exposes you to radiation. You should talk to your doctor about the benefits and drawbacks of mammograms. Together, you can decide when to start and how often to have a mammogram.

Now while it's difficult to figure out for physicians by seeing only images of x-ray that weather the tumor is toxic or not training a machine learning model according to the identification of tumor can be of great help.

PROBLEM STATEMENT & DISCUSSION

Breast Cancer is one of the leading cancers developed in many countries including India. Though the endurance rate is high – with early diagnosis 97% women can survive for more than 5 years. Statistically, the death toll due to this disease has increased drastically in last few decades. The main issue pertaining to its cure is early recognition. Hence, apart from medicinal solutions some Data Science solution needs to be integrated for resolving the death causing issue.

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.

PROPOSED SOLUTION & RESULT ANALYSIS

In this project we will use Data Mining and Machine Learning Algorithms to detect breast cancer, based off of data. Breast Cancer (BC) is a common cancer for women around the world.

Early detection of BC can greatly improve prognosis and survival chances by promoting clinical treatment to patients.

We will use the UCI Machine Learning Repository for breast cancer dataset.

Url: http://archive.ics.uci.edu/ml/datasets/breast+cancer+%20wisconsin+%28diagnostic%29

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. To create the dataset Dr. Wolberg used fluid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which is capable of perform the analysis of cytological features based on a digital scan. The program uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, then it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector Attribute Information:

1. ID number 2) Diagnosis (M = malignant, B = benign) 3–32)

[Ten real-valued features are computed for each cell nucleus:]

2. radius (mean of distances from center to points on the perimeter)
3. texture (standard deviation of gray-scale values)
4. perimeter
5. area
6. smoothness (local variation in radius lengths)
7. compactness (perimeter² / area — 1.0)
8. concavity (severity of concave portions of the contour)
9. concave points (number of concave portions of the contour)
10. symmetry
11. fractal dimension (“coastline approximation” — 1)

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 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

DATA MINING AND MACHINE LEARNING

The term "data mining" is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence (e.g., machine learning) and business intelligence. The book Data mining: Practical machine learning tools and techniques with Java [8] (which covers mostly machine learning material) was originally to be named just Practical machine learning, and the term data mining was only added for marketing reasons. Often the more general terms (large scale) data analysis and analytics – or, when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

In this project we use the following machine learning algorithms:

Decision tree algorithms:

Decision tree algorithms are successful machine learning classification techniques. They are the supervised learning methods which use information gained and pruned to improve results. Moreover, decision tree algorithms are commonly used for classification in many research, for example, in the medicine area and health issues. There are many kinds of decision tree algorithms such as ID3 and C4.5. However, J48 is the most popular decision tree algorithm. J48 is the implementation of an improved version of C4.5 and isan extension of ID3.

K-nearest-neighbor's (kNN) algorithm:

It is a simple supervised learning algorithm in pattern recognition. It is one of the most popular neighborhood classifiers due to its simplicity and efficiency in the field of machine learning. KNN algorithm stores all cases and classifies new cases based on similarity measures; it searches the pattern space for the k training tuples that are closest to the unknown tuples. The performance depends on the optimal number of neighbors (k) chosen, which is different from one data sample to another.

Support Vector Machine (SVM):

It is a supervised learning method derived from statistical learning theory for the classification of both linear and nonlinear data. SVM classifies data into two classes over a hyperplane at the same time avoiding over-fitting the data by maximizing the margin of hyperplane separating.

Naïve Bayes (NB) It is a probabilistic classifier:

It is one of the most efficient classification algorithms based on applying Bayes' theorem with strong (naïve) independent assumptions. It assumes the value of the feature is independent of the value of any other features, given the class variable.

Based on the maximum probability. It detects the class membership for the given tuple to a particular class.

Logistic regression:

In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one. Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable (target) is categorical.

DATA EXPLORATION

We will first go with importing the necessary libraries and import our dataset to colab.research.google.com.

We can examine the data set using the pandas' head () method.

df.head(7) {first 7 rows of the data}

We can find the dimensions of the data set using the panda dataset 'shape' attribute.

We can observe that the data set contain 569 rows and 33 columns. 'Diagnosis' is the column which we are going to predict, which says if the cancer is M = malignant or B = benign. 1 means the cancer is malignant and 0 means benign. We can identify that out of the 569 persons, 357 are labeled as B (benign) and 212 as M (malignant).

Each row represents a patient and 33 features on the 569 patients.

The last column Unnamed 32 has NaN values so we need to remove that column with empty values.

So, we count the number of empty columns and drop the columns with empty values.

So, column Unnamed: 32 has 569 missing values so we drop it.

So, the new shape of the data is (569,32) which means 569 rows and 32 columns

We can see that id column acts as the identifier of the patient and it is of integer type and it cannot be used as a feature to predict the tumor

Next, we encode categorical data values

Here the value 1 represents Malignant (M) (harmful) and value 0 represents benign (B) cells (not harmful) cells.

Now we visualize a correlation between the different attributes.

In this heat map we can see how much one column influences all the other columns (e.g., radius mean has 32% influence on texture mean).

TRAINING AND TESTING

Next, we split the datasets into independent (X) and dependent (Y) data sets.

They are of type array.

The dependent data set (Y) has the diagnosis whether the patient has cancer and the independent data set (X) has the features that are used to predict the outcome.

Now we split the data set into 75% training and 25% testing and use different machine learning models such as logistic regression, random forest classifier, decision tree to the training data.

So, we can see that the decision tree classifier kas the best accuracy among all the models i.e., 100%

We will now predict the test set results and check the accuracy with each of our model.

To check the accuracy, we need to import confusion matrix method of method of metrics class. The confusion matrix is way of tabulating the number of mis-classifications i.e., the number of predicted classes which ended up in a wrong classification in based on the true classes.

Here the matrices are of form

[TP FP]

[FN TP]

were

TP is true positive: A true positive is an outcome where the model correctly predicts the positive class
TN is true negative: A true negative is an outcome where the model correctly predicts the negative class.

FN is false negative: A false negative is an outcome where the model incorrectly predicts the negative class.

FP is false positive: A false positive is an outcome where the model incorrectly predicts the positive class.

Based on the test data we can see that Model 5 ie Random forest classifier has 96.5% accuracy on the test data so we can use it to predict the actual outcome whether a patient has cancer or not.

PREDICTION OF MODEL

So here we printed the predictions. The first data shows the actual result of which patient had cancer and the second data is the one predicted by the model.

The accuracy of the model is 96.5% so we can see a few wrong predictions but mostly this model is successful in predicting a tumor Malignant (M) (harmful) or Benign (B) (not harmful) based upon the features provided in the data and the training given.

CONCLUSION & FUTURE SCOPE

In this project in python, we learned to build a breast cancer tumor predictor on the Wisconsin dataset and created graphs and results for the same. It has been observed that a good dataset provides better accuracy. Selection of appropriate algorithms with good home dataset will lead to the development of prediction systems. These systems can assist in proper treatment methods for a patient diagnosed with breast cancer. There are many treatments for a patient based on breast cancer stage; data mining and machine learning can be a very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable databases.

Done by

- ZAMEEL ALI MOHAMMED
- PARAS KUSHWAHA
- SRIDHAR MALLARAPU
- SULGANA MAJI