GRAPHIC ERA HILL UNIVERSITY

DEHRADUN

A Project Report

On

BREAST CANCER DETECTION

USING

LOGISTIC REGRESSION

By **Yogesh Kumar Bhatt**

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

(2022-2023)

DECLARATION

I Yogesh Kumar Bhatt, student of B-Tech Computer Science Engineering, Graphic Era Hill University Dehradun declare that the technical project work entitled “Breast Cancer Detection Using Logistic Regression” has been carried out by me and submitted in partial fulfilment of the course requirements for the award of degree in B-Tech of Graphic Era Hill University during the academic year 2022-2023. The matter embodied in this synopsis has not been submitted to any other institution for the award of any other degree.

Date:

Yogesh Kumar Bhatt

Student ID: 20011189

Section: H

Semester: 5th

CERTIFICATE

This is to certify that the project report entitled “Breast Cancer Detection Using Logistic Regression” is a Bonafide project work carried out by Yogesh Kumar Bhatt, Roll No- 2018886, in partial fulfilment of award of Degree of B-Tech of Graphic Era Hill University, Dehradun during the Academic Year 2022-2023. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated. The project has been approved as it satisfies the academic requirements associated with the degree mentioned.

ACKNOWLEDGEMENT

Hereby, I am submitting the project report on “Breast Cancer Detection Using Logistic Regression” as per the scheme of Graphic Era Hill University, Dehradun. I consider it mine cardinal duty to express the deepest sense of gratitude to Mr. Deepak Sing Rana, Asst. Professor, Department of Computer Science and Technology for the invaluable guidance extended at every stage and in every feasible way.

Finally, I am very thankful to all the faculty members of the Department of Computer Science and Technology, friends and my parents for their constant encouragement, support and help throughout the period of project conduction.

Name: Yogesh Kumar Bhatt

Student ID: 20011189

University Roll No.: 2018886

Section: H

Semester: 5th

Session: 2022-2023

Introduction

Breast cancer is now one of the most prevailing cancers that affects humans, especially woman, and early diagnosis would go a long way to reducing the damage done by this cancer on its victims. Breast cancer's causes are multifactorial and involve family history, obesity, hormones, radiation therapy, and even reproductive factors. Every year, one million women are newly diagnosed with breast cancer, according to the report of the world health organization half of them would die, because
it's usually late when doctors detect the cancer (Aaltonen et al., 1998). Breast cancer can be categorized into two, which are malignant breast cancer and benign breast cancer. The classification of breast cancer as either malignant or benign is possible by scientifically studying the features of breast tumours, lumps, or any abnormalities found in the breast. At the benign stage the cancer has less risk and is not lifethreatening while cancer that is categorized as malignant is life-threatening (Huang, Chen, Lin, Ke, & Tsai, 2017). Malignant tumours expand to the neighbouring cells, which can spread to other parts, whereas benign masses can't expand to other tissues, the expansion is then only limited to the benign mass (Aaltonen et al., 1998; To accurately classify breast cancer as benign or malignant, researchers have employed an aspect of Artificial intelligence (AI) which is machine learning. Machine learning algorithms are used to build models that accept as input, attributes that qualify a breast cancer case and produce as output a label for the type of the cancer, label 1 for being benign or label 2 for malignant.

Classification

Classification in data mining involves basically two processes: firstly it is the model training and with a test data to determine the class label of unknown test instances; secondly is the performance evaluation to check the accuracy of the classifier model, that is calculating the differences between the classified and actual values for each attribute tuple in the test dataset.

Basic Terminologies and Concepts

Machine Learning (ML) is the science (and art) of programming computers so they can learn from data (Géron, 2017). Machine learning can be defined in a more general way as: ML as the field of study that gives computers the ability to learn without being explicitly programmed. – Arthur Samuel, 1959. ML can also be defined in a more technical way as: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T as measured by P, improves with experience E. – Tom Mitchell,1997.

There are several applications for ML, the most significant of which is data mining.

People are often prone to making mistakes during analyses or, possibly, when trying to establish relationships between multiple features.

Supervised Learning

Supervised machine learning is the search for algorithms that cogitate from externally supplied instances to give general hypotheses, which then infer predictions about future instances. In other words, the goal of supervised learning is to build an incisive model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown.

Logistic regression model

Logistic regression was developed in late the 1960s and early 1970s(Cabrera, 2007; Haigh, Cox, & Snell, 2007; Peng, Lee, & Ingersoll, 2002) and became popular among researches in various fields, particularly among health researchers(Abedin, Chowdhury, & Afzal, 2016).

Logistic regression is prevalent in almost every standard statistical software package.

Owing to its wide popularity and usefulness in research it is important to comprehend the basics of logistic regression i.e., how does the model operate, what postulations are needed to be verified, how to report the results found, etc. (Abedin et al., 2016). In this paper, we study binomial logistic regression for the classification of breast cancer dataset. The mathematical notion of logistic regression is to show the relationship between the outcome variable (dependent variable) and predictor variables (independent variables) in terms of logit: the natural logarithm of odds. Let’s take into consideration a simple case where Y is a dichotomous dependent variable categorized as “1” and “0”

and X is a continuous independent variable. Now if we draw a scatter plot, as expected we will have two parallel lines analogous to each dependent variable category.

Objective

The aim of this analysis is to use Logistic Regression to classify the data into two classes of diagnosis— Malignant & Benign. The evaluation metrics used are accuracy, ROC, confusion matrix, precision-recall.

Dataset

The Wisconsin Breast Cancer (Diagnostic) dataset has been extracted from the UCI Machine Learning Repository. 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.

Class distribution: 357 benign, 212 malignant

Number of instances: 569; Number of attributes: 32

Attributes

Ten real-valued features are computed for each cell nucleus

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area — 1.0)

g) concavity (severity of concave portions of the contour)

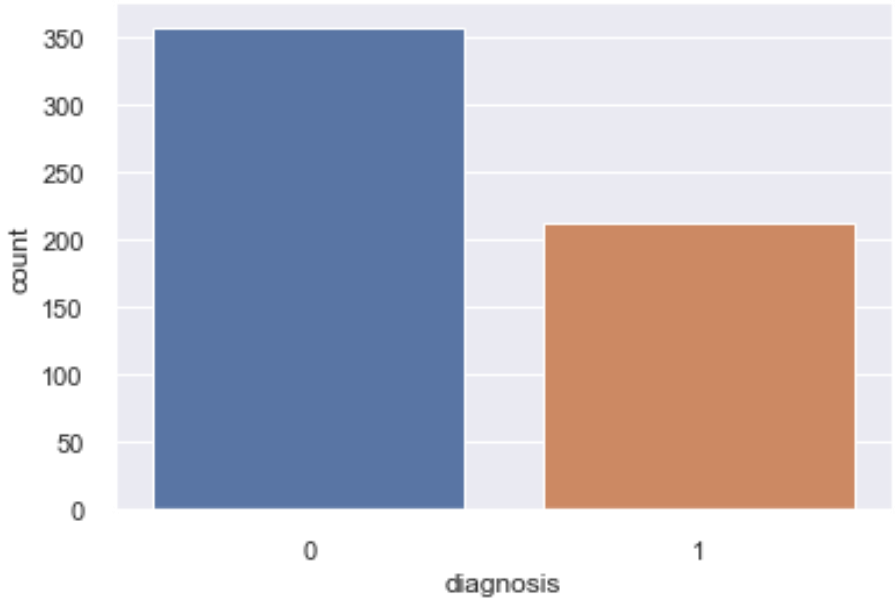
h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension (“coastline approximation” — 1)

Preprocessing

The dataset provided is already clean and does not have any missing values. As a part of the preprocessing stage, the data is standardized using Standard Scaler library. It transforms the attributes to normal distribution.



Evaluation metrics for classification

Accuracy cannot be completely relied upon as an evaluation metric for Classification.

Confusion Matrix: It is a matrix representation of the results of any binary testing. In this study, the hypothesis is to validate if the person having the lump is malignant or benign. Confusion matrix is used to represent the actual values and the prediction values.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Confusion Matrix:

True positive (TP): Malignant lump correctly identified as malignant

True negative (TN): Benign lump correctly identified as benign

False positive (FP): Benign lump incorrectly identified as malignant

False negative (FN): Malignant lump incorrectly identified as benign

- Accuracy: Accuracy is the proportion of true results among the total number of cases sample: (TP +TN) / (TP + TN + FP +FN)
- Precision: The proportion of predicted Positives that is actually positive: TP / (TP + FP)
- Recall OR Sensitivity OR TPR: The proportion of actual positives correctly classified: TP / (TP + FN)
- Specificity: The proportion of actual negatives correctly classified: TN/(TN + FP)
- F1 Score: F1 score kind of maintains a balance between the precision and recall for the classifier by giving equal importance to both precision and recall. Lower the precision, lower is the F1 score. Lower the recall, lower is the F1 score. The F1 value is between 0 and 1 and is the mean of precision and recall: 2 * (Precision * Recall / Precision + Recall)

AUC ROC: Area Under Curve (AUC) is the area covered under the ROC. Receiver Operating Characteristics. ROC is the ratio of True Positive Rate (TPR) and False Positive Rate (FPR) where TPR is proportion of actual positives correctly classified by the model and FPR is the proportion of false positives classified by the model. We have the above values (TPR, FPR) from the model. We can use various threshold values between 0.1–1 to plot the sensitivity and 1-Specificity on the graph. The curve derived is the ROC curve. Lower threshold can increase the number of FP and decrease the number of FN. Higher threshold can decrease the number of FP and increase the FN.

Classification using Logistic Regression (Using RFE for feature elimination)

After splitting the data into training and test set, the training data is fit and predicted using Logistic Regression with GridSearchCV. GridSearchCV is a function that belongs to the sklearn library. It is a process to optimize the model performance by tuning the hyperparameters, automatically, by setting a few parameters in the function. Choosing optimal features for model training is an extremely difficult task to achieve. Feature selection is a key technique to improve the performance of any model. Feature selection can be done using Principal Component Analysis, but it provides us with the principal components which are achieved by a combination of features. It does not clearly declare the features, used to calculate the principal components.

Recursive Feature Elimination technique selects individual features by recursively training the data for a given model. Thus, a set of features with high importance is used to improve performance of the model.