Nitte Meenakshi Institute of Technology,

Department of Computer Science and Engineering

18CSE751 Introduction to Machine Learning

Learning Activity Proposal

Breast Cancer Detection

Prity Panjiyar(1NT18CS215), Kirti Chaudhary(1NT18CS197)

Abstract:

Breast Cancer is a common cancer found in women around the world, and early detection of Breast Cancer can greatly improve prognosis and survival chances by promoting early treatment to patients .This proposal focuses on development of a tool that will help to detect Breast Cancer using data, python and machine learning.

Introduction:

Machine Learning refers to a system's ability to acquire and integrate knowledge through large-scale observations to improve and extend itself by learning new knowledge rather than by being programmed with that knowledge.

The World Health Organization (WHO) reported that breast cancer has posed a threat to approximately 2.1 million women in the world every year. Additionally, it stated that 627,000 women lost their life from breast cancer in 2018.

This proposal focuses on development of a tool that will help to detect Breast Cancer using data, python and machine learning. Machine Learning models like Logistic Regression, Decision Tree Classifier, Random Forest Classifier will be used to develop this tool.

Data Set:

We will be using data.csv (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data) data set from kaggle.
The features in this data set are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of a cell nuclei present in the image. Each row of data represents a patient that may or may not have cancer.

Attribute Information:

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

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

a) radius (mean of distances from centre to points on the perimeter)
b) texture (standard deviation of grey-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)

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.

All feature values are recoded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

Machine Learning Methods:

The following Machine Learning methods will be used:-

1. Logistic Regression:

As logistic regression predicts the output of a categorical dependent variable. So, with the help of this model we will train our model to determine whether the person has cancer or not.

2. Decision Tree Classifier:

With the help of decision tree classifier we will represent the features of our data set, represent the decision rules and also the outcome.

3. Random Forest Classifier:

With the help of random forest classifier we will improve the predictive accuracy of our data set.

Assessment:

Train/test split:

We will be using Train/test split to validate our models. In this method we simply split our data randomly into roughly 75% used for training the model and 25% for testing the model.
The advantage of this approach is that we can see how the model reacts to previously unseen data.

Presentation and Visualisation:

We will be using python as our programming language and google colab as our platform as all the python libraries can be imported easily in google colab. And the output will be in the form of simple tables and graphs.

Roles:

Prity:- Coding, Literature survey, Making reports

Kirti:- Coding, Literature survey

Schedule:

Date	Task to be Completed
15/12/2021	Selected the project
16/12/2021	Began literature survey
20/12/2021	Completed the proposal

Bibliography:

1. Bray et al. (2018) Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6):394–424, 2018.
2. Broeders et al. (2012) Mireille Broeders, Sue Moss, Lennarth Nyström, Sisse Njor, Håkan Jonsson, Ellen Paap, Nathalie Massat, Stephen Duffy, Elsebeth Lynge, and Eugenio Paci. The impact of mammographic screening on breast cancer mortality in europe: a review of observational studies. Journal of medical screening, 19(1_suppl):14–25, 2012.

3. Dibisa et al. (2019) Teshale Mulatu Dibisa, Tilayie Feto Gelano, Lemma Negesa, Tewelde Gebre Hawareya, and Degu Abate. Breast cancer screening practice and its associated factors among women in kersa district, eastern ethiopia. The Pan African Medical Journal, 33, 2019
4. Hadgu et al. (2018) Endale Hadgu, Daniel Seifu, Wondemagegnhu Tigneh, Yonas Bokretsion, Abebe Bekele, Markos Abebe, Thomas Sollie, Sofia D Merajver, Christina Karlsson, and Mats G Karlsson. Breast cancer in ethiopia: evidence for geographic difference in the distribution of molecular subtypes in africa. BMC women's health, 18(1):40, 2018.