Breast Cancer Prediction
Milestone: Project Proposal

Group 25

Student 1: Siva Vasantha Harika Mangu
Student 2: Raaga Sindhu Mangalagiri

mangu.s@northeastern.edu
mangalagiri.r@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Siva Vasantha Harika Mangu

Signature of Student 2: Raaga Sindhu Mangalagiri

Submission Date: 02/03/2023

Project Proposal

IE 7275: Data Mining in Engineering

Problem Statement:

About 1 in 8 women will develop invasive breast cancer over the course of their lifetime. And every year around 40,000 women in United States alone, are dying from breast cancer. It starts when cells in breast begin to grow out of control. These cells usually form tumors that can be seen via X-ray often felt as lumps in the breast area. And one of the shocking revelations can be said that it is found in women who don't show any symptoms. It is mostly occurring disease among women and in specific elderly age groups. They cannot prevent or control their risk of cancer, and it cannot even be recognized at early stage due to lack of symptoms and abnormalities they face. Hence, by building a machine learning model, we can be able to predict the risks of breast cancer among women at early age, based on different attributes acting as a contributing factor to the disease.

Problem Definition:

Given a set of patient data including demographic information, medical history, and medical imaging results, the goal is to develop a model or system that can accurately predict the likelihood of a patient developing breast cancer. The challenge is to classify these tumors into malignant(cancerous) or benign(non cancerous). The model should be able to handle missing or incomplete data and be able to generalize well to new unseen cases. Additionally, the model should be interpretable and provide insights on the most important features/variables that contribute to the prediction. The goal is to improve the early detection and prevention of breast cancer by identifying high-risk individuals and providing them with the necessary interventions, resulting in better outcomes for patients.

Data Sources:

Kaggle

Breast Cancer Dataset | Kaggle

Data Description:

Breast cancer data set consists of 32 attributes and 570 records in total, in which 30 attributes are contributing as predictors and the attribute diagnosis is the response variable which predicts whether the patient is Benign or malignant. The Unique ID is the primary key which is the not the predictor either or a response variable. The predictors such as radius_mean,texture_mean,perimeter_mean,area_mean,smoothness_mean,compactness_mean,concavity_mean,concavepoints_mean,symmetry_mean,fractal_dimension_mean,radius_se,texture_se,perimeter_se,area_se,smoothness_se,compactness_worst,concavity_worst,concavepoints_worst,symmetry_worst,fractal_dimension_worst.Some of the other risk factors which can be considered as predictors are age, family history, certain genetic mutations, and certain lifestyle factors such as alcohol consumption and lack of physical activity. Symptoms of breast cancer include a lump or thickening in the breast tissue, changes in the size or shape of the breast, and changes to the skin on the breast such as redness or dimpling.