Overview

Breast cancer is the most commonly diagnosed cancer in women in the United States. Despite the low mortality rate of breast cancer, it is desired to detect and introduce intervention in the early stage to prevent further complications. The most effective method for screening breast cancer is mammography, which involves taking an x-ray image of the breast. Breast biopsy is also used in addition to mammography on suspicious lesion but is often an unpleasant process. Therefore, there has been a high demand in utilizing machine learning on mammography data to assist diagnosis. In this project, I will be working with mammography data and develop a logistic regression model to classify a mammographic mass lesion as either benign or malignant.

Research Question

The inferential goal of this research project is to use mammography data to determine the effects that certain attributes have on the odds of developing a malignant lesion. Specifically:

- Does certain mass shape lead to higher odds of a malignant lesion?
- How does mass density affect the severity of a lesion?
- What attribute is the most critical predictor at determining the odds of a malignant lesion?

Data

The data that I will use is the Mammography Mass Data Set sourced from the UCI Machine Learning Repository. This dataset contains 961 instances and 6 attributes including the response. There are 445 positive cases and 516 negative cases. The dataset contains some missing values, which will be imputated using multiple imputation. The attribute detail is summarized below:

Name	description	Type	Missing
BI-RADS	BI-RADS assessment ranging from 1 (definitely benign) to 5 (highly suggestive of malignancy)	ordinal	5
Age	Patient's age	integer	5
Shape	mass shape: round=1, oval=2, lobular=3, irregular=4	nominal	31
Margin	mass margin: circumscribed=1, microlobulated=2, obscured=3, ill-defined=4, spiculated=5	nominal	48
Density	mass density high=1 iso=2 low=3 fat-containing=4	ordinal	76
Severity	Response: benign=0 or malignant=1	Binary	0

Project Plan

The model that I will use is logistic regression model and the model will be developed in R. The timeline is summarized below:

Date	Deliverables
Oct 21st – Oct 29th	Perform EDA (interaction), Imputation of missing values
Oct 30th – Nov 5th	Perform model selection and model assessment
Oct 6th – Nov 8th	Prepare and record presentation
Nov 8th – Nov 15th	Finish project report