Predicting Breast Cancer Project Proposal

Jonathan Daughtry

February 2021

Overview

The problem that is being addressed in this project is the potential to recognize and predict the likelihood of a tumor being malignant in patients with breast tumors. Thankfully, research and treatments have greatly improved for breast cancer, but one of the key indicators of success is early detection. Cancer has affected almost everyone in this world, whether directly or through a relative or friend, and breast cancer is one of the most common types of cancer in women, so this is an issue that everyone can agree is very important. The stakeholders that would find this the most useful were doctors that were trying to predict the likelihood of a tumor being malignant and the patients involved.

Data

The dataset that is going to be used for this project is the Breast Cancer Wisconsin dataset from kaggle. This dataset has characteristics of 357 benign and 212 malignant tumors. The dataset focuses on the characteristics of the nucleus of the tumor cells, including the radius, texture, perimeter, smoothness, etc. The goal is to determine which of the characteristics that are given are the most important in determining if a tumor is benign.

Approach

Once the data is explored, models will be used to determine which the best at predicting the outcome of the tumor.

I anticipate using several modeling algorithms to determine the best outcome for this regression model. The models that I would like to use areLogistic Regression, K-Nearest Neighbors, and Random Forest. I am also leaving the possibility of using Naive Bayes Classifier and Support Vector Machines. I also anticipate using Tableau to present the information to those who might be interested in the information. Tableau will be helpful in visualizing the different confusion matrices produced, the F1 Scores, and how the AUC changes as the prediction parameters change. Gridsearch cross validation will be used to attempt to find the best parameters for each model. Using the best model available, the doctor will be able to not only predict whether the tumor is malignant or benign, but also will be able to give a reasonable estimate on the likelihood that the prediction is accurate.

Deliverables

- The code that I developed throughout each part of the machine learning process through Jupyter Notebook, which includes the data cleaning, EDA, and the modeling
- The visuals for the EDA and the modeling results through Tableau so they can be further explored
- Final presentation through slides deck
- Detailed report of the project, including a description of each step taken