[[1]](#footnote-1)

Cancer Prediction – Benign or Malignant(March 2019)

Martin Garcia, marting@smu.edu, Andrew L. Wilkins, awilkins@mail.smu.edu

*Abstract*—Data was collected from the UCI Machine Learning Respository for Breast Cancer. The data includes multiple explanatory variables of the cell being observed and a binary response variable of malignant or benign.

*Index Terms*—Cancer, Breast Cancer, malignant, benign, classification model, healthcare, supervised classification, discriminant analysis, principal component analysis, binary response

# Overview

T

HE motivation for this project is to help predict if a cancer is malignant or benign based on 30 different features.

# Data Description

The data includes characteristics of cell nuclei. Response variable is a binary and categorical with 1 as malignant and 0 as benign. Explanatory variables include measurements of cells of data type numeric and continuous.

|  |  |
| --- | --- |
| id | smoothness\_se |
| diagnosis (response) | compactness\_se |
| radius\_mean | concavity\_se |
| texture\_mean | concave points\_se |
| perimeter\_mean | symmetry\_se |
| area\_mean | fractal\_dimension\_se |
| smoothness\_mean | radius\_worst |
| compactness\_mean | texture\_worst |
| concavity\_mean | perimeter\_worst |
| concave points\_mean | area\_worst |
| symmetry\_mean | smoothness\_worst |
| fractal\_dimension\_mean | compactness\_worst |
| radius\_se | concavity\_worst |
| texture\_se | concave points\_worst |
| perimeter\_se | symmetry\_worst |
| area\_se | fractal\_dimension\_worst |

# Analysis

To begin our analysis we will start with our exploratory data analysis to get a higher level understanding of our variables. This will help us see if transformations are required by looking at variance, distributions, linearity, and sample size requirements. Once our data is transformed we will remove variables that express multicollinearity with principal component analysis. We will use discriminant analysis to classify the cells as either malignant or benign type. This means creating a cross validation model with training and test set to help create our model. We can also use a MANOVA to help see the impact of individual explanatory variables to our response. This also means accounting for a binary response which would incorporate sample proportions. A simple linear regression model in matrix form can also help us to see which variables give us the best explanation in variation.

# Goal

Overall goal will be to create a predictive model that will help us classify cells as either malignant or benign. We also wish to see which explanatory variables contribute to helping us make our prediction.

# RESOURCES NEEDED

In order to complete this project we will utilize SAS for statistics measures and python for organizing our data to create visualizations reflective of our findings.

# REFERENCES

[1] <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/activity>

[2] <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>

1. [↑](#footnote-ref-1)