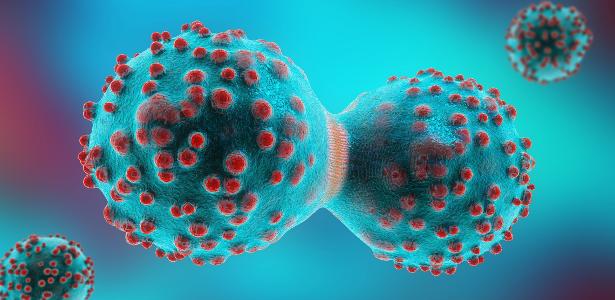
MSDS 6372 Project 2: Breast Cancer Analysis

Jaime Villanueva, Aaron Faltesek, Luay Dajani

August 5, 2018

# Introduction

 Image 1: Cancer Cells Dividing (ref 1)

DATA SCIENCE & STATISTICS PROJECT: breast cancer dataset

## Purpose

This is a training exercise using a well-known and circulated data set to discover means for using data science and statistical methods and models to determine the diagnosis for breast cancer in included patients. Though out of the scope of this excersise, creating accurate models which accurately predict the state of a tumor (cencerous or not) may help to reduce time, invasivness, and money to be able to more quickly identify or flag any problems, ultimately saving lives.

## Background

Breast cancer is a cancer that develops in the breast tissue and may appear in as a lump in the tissue, with a different consistently that the rest of the tissue, shape change in the breast, or by other means (ref 2). In many cases, breast cancer can be easily diagnosed via a biopsy, though others may require higher level lab analysis (ref 3). In this excersise, different sugical biopsy measurements (ref 4) will be examined through different methods to see which model most accurately reflects, as well as measurements (variables) tend to best contribute to determining if a tumore is most likely cancerous or not.

## Methodology

This excersise will demonstrate three models which include logistic regression, Linear/Quadratic Discriminiation Analysis, and Random Forest.

# Data Set Description

The breast cancer data set comprises of 699 total observations with 11 total variables or readings based on a well-known breast cancer dataset based on research from Dr. Wolberg at the University of Wisconsin facilities (ref 4). The readings were collected over a 2 years (1989-1991) and released in 1992 as follows:

Group 1: 367 instances (January 1989)  
Group 2: 70 instances (October 1989)  
Group 3: 31 instances (February 1990)  
Group 4: 17 instances (April 1990)  
Group 5: 48 instances (August 1990)  
Group 6: 49 instances (Updated January 1991)  
Group 7: 31 instances (June 1991)  
Group 8: 86 instances (November 1991)  
-----------------------------------------  
Total: 699 points (as of the donated datbase on 15 July 1992)

Each variable has a “normalized” or “Standardized” ranking from 1 to 10 (1 being most minor and 10 being the most major) based on the the measurement or severity of the attribute compared with other readings in the same category except for the sample code and the outcome. The readings were stardarized for simpler calculations and usage in the models. These variables include:

|  |  |  |
| --- | --- | --- |
| Attribute | Full Name | Description |
| sample\_code | Sample Code | Unique Observation Number |
| clump\_thick | Clump Thickness | Reletive thickness/desnity of the tumor |
| unif\_cellsize | Uniformity of Cell Size | ??? |
| unif\_cellshape | ??? | ??? |
| marg\_adhes | ??? | ??? |
| epi\_cell\_size | ??? | ??? |
| bare\_nuclei | ??? | ??? |
| bland\_chro | ??? | ??? |
| normal\_nucleoli | ??? | ??? |
| mitosis | ??? | ??? |
| Outcome | ??? | ??? |

# Exploratory Data Analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sample\_code | clump\_thick | unif\_cellsize | unit\_cellshape | marg\_adhes |
| 1000025 | 5 | 1 | 1 | 1 |
| 1002945 | 5 | 4 | 4 | 5 |
| 1015425 | 3 | 1 | 1 | 1 |
| 1016277 | 6 | 8 | 8 | 1 |
| 1017023 | 4 | 1 | 1 | 3 |
| 1017122 | 8 | 10 | 10 | 8 |

## sample\_code clump\_thick unif\_cellsize unit\_cellshape marg\_adhes  
## 1 1000025 5 1 1 1  
## 2 1002945 5 4 4 5  
## 3 1015425 3 1 1 1  
## 4 1016277 6 8 8 1  
## 5 1017023 4 1 1 3  
## 6 1017122 8 10 10 8

## References

#### 1: Image 1: <https://conteudo.imguol.com.br/c/entretenimento/30/2017/12/20/divisao-celula-cancer-1513772032641_v2_615x300.jpg>

#### 2: Wikipedia (1st paragraph): Breast cancer: <https://en.wikipedia.org/wiki/Breast_cancer>

#### 3: Wikipedia (Diagnosis): <https://en.wikipedia.org/wiki/Breast_cancer#Diagnosis>

#### 4: Computerized breast cancer diagnosis and prognosis from fine-needle aspirates: <https://www.researchgate.net/publication/15451281_Computerized_breast_cancer_diagnosis_and_prognosis_from_fine-needle_aspirates>