Fundamental of Machine Learning

Final Project

Linear and Logistics regression predicting the radius and malignancy of breast cancer tumors

KENT STATE UNIVERSITY
OHIO · 1910

Nawwaf Albahar

Fall 2022

Kent State University

Introduction

Breast cancer is the leading type of cancer worldwide among women. Even though it can occur in both sexes, it is more than 100 times more common in women than in men. In 2018 alone, it accounted for 2 million cases and 627 thousands deaths. (Mayo clinic, 2019)

Breast disorders can be either noncancerous (benign) or cancerous (malignant). Most cases are benign, and consequently do not threaten the life of a patient. However, more severe malignant cases can lead to a mastectomy, removal of breast tissue, or even death.

Although various hormonal, lifestyle and environmental factors have been identified to increase the risk of developing breast cancer, it is not entirely clear to researchers why certain people get it and others do not. Early detection of the cancer is often vital for the treatment to be successful. Therefore, frequent self-examinations and medical screenings are of great importance to further increase survival rates. (Centers for Disease Control and Prevention. n.d)

Nonetheless, the key to the right allocation of treatment is correctly distinguishing between cancerous and noncancerous cases, which sometimes proves to be problematic. This is why I decided to create a Decision Support System (DSS) which can help doctors correctly differentiate between benign and malignant cases of cancer.

The dataset used for the development of the tool is Breast Cancer Wisconsin (Diagnostic) Data Set. It consists of 569 cases of breast disorders with features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass, with approximately. 40% of cases being malignant. Features “describe characteristics of the cell nuclei” and are as follows (name of the variable after equation sign):

“a) radius (mean of distances from center to points on the perimeter) = radius_mean
b) texture (standard deviation of gray-scale values) = texture_mean
c) perimeter = perimeter_mean
d) area = area_mean
e) smoothness (local variation in radius lengths) = smoothness_mean
f) compactness (perimeter^2 / area - 1.0) = compactness_mean
g) concavity (severity of concave portions of the contour) = concavity_mean
h) concave points (number of concave portions of the contour) = concavity_mean
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" (or largest which is the mean of the three largest values) of these features are computed for each image, resulting in 30 features.” (Kaggle.com, 2017)

Aside from the features, the dataset also contains a classification of the diagnosis of the breast disorder as benign or malignant.

The decision support system I created is built on multiple models. Firstly, there is a standard linear regression model that predicts radius (mean of distances from center to points on the perimeter) of the breast disorder. Then, a logistic regression model predicts the likelihood of a patient to have either a benign or malignant tumor.

The target user of the DSS are doctors, especially oncologists, working in a hospital who are responsible for detecting breast cancer and recommending treatment. The tool helps them distinguish between benign and malignant cases by analyzing different features of the tumor, which aids the choice of the right treatment for the patient. The tool should be used alongside conducting an interview with the patient, as considering lifestyle and environmental factors can also aid the doctor in determining the likelihood of the cancer being malignant.
Specifically, I expect doctors to use the system to decide whether a biopsy is necessary or not.

The next section, “Machine Learning Techniques used in the DSS”, clarifies the models used and justifies their helpfulness in achieving the goals of the DSS. Lastly, the DSS is critically reflected on.

Machine Learning Techniques Used

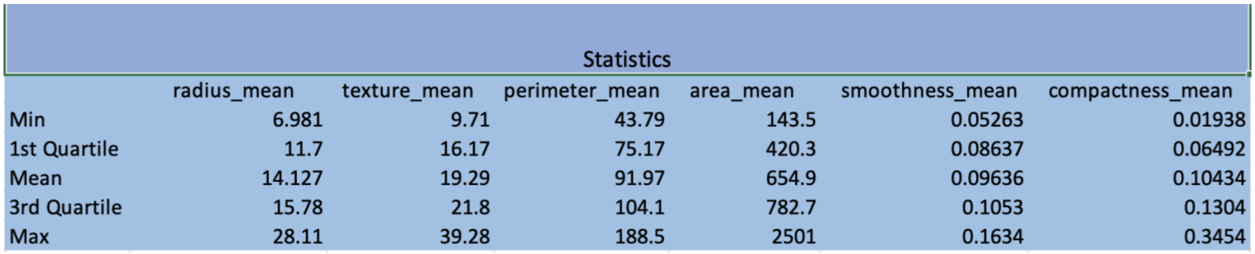
Linear Regression

Introduction

The purpose of this summary is to implement and test several standard linear regression models to develop a radius mean prediction model for Breast Cancer Wisconsin. The columns of the data include texture mean, perimeter mean, area mean, smoothness mean, compactness mean, concavity mean, concave points mean, symmetry mean and fractal dimension mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se etc.nThe objective of this machine learning method is to predict the most possible accurate radius mean of the Patients in Breast Cancer Wisconsin dataset using the radius mean of patients’ tumors.

Data loading and splitting

As a first step I load the Breast Cancer Wisconsin dataset that I acquired from Kaggle, and I examine the summary of the dataset. The below aggregations show us the average, minimum and maximum values in the dataset. This allows us to quickly see if data cleaning is needed (for example one value has impossibly high or low value).



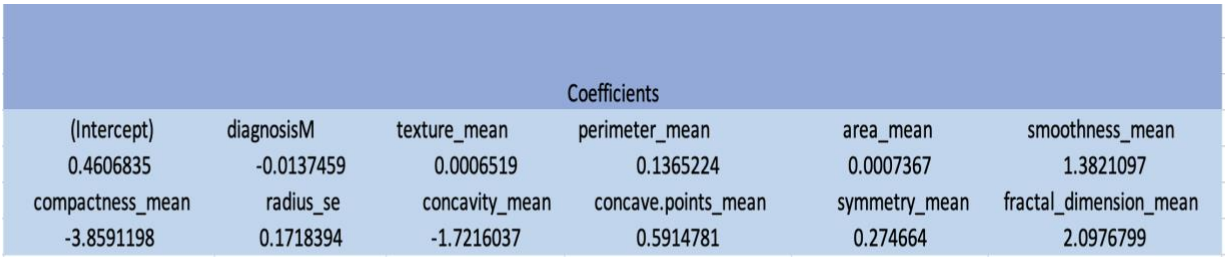
Statistics

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
Min	6.981	9.71	43.79	143.5	0.05263	0.01938
1st Quartile	11.7	16.17	75.17	420.3	0.08637	0.06492
Mean	14.127	19.29	91.97	654.9	0.09636	0.10434
3rd Quartile	15.78	21.8	104.1	782.7	0.1053	0.1304
Max	28.11	39.28	188.5	2501	0.1634	0.3454

After looking at the summary, the variables look within reasonable range. As I can observe, there are no obviously outlying values in the above summary, which indicates that there is no need to clean the data, it is already in a processable state. After examining the data, I split it into two disjoint sets: training and evaluating sets. The sample() function I use selects randomly the rows to be put in either sets. I make a 60-40 split, which means 60% of rows are put into the training set and the rest is into the evaluation (out of 569 rows, 341 is for training and 228 is for evaluating).

Implementation of Linear Regression

I implemented a standard linear regression model for predicting the radius_means. The patients radius_means in Breast Cancer Wisconsin dataset are impacted by several distinct factors, which are different in their impact on the radius_mean level. Therefore, I examine the impact of each variable on the target value by examining the standard error rate of each variable. The larger the number is, the more impact a single unit change has on the tumor radius_mean. If the number is negative, it influences negatively the radius_mean, and a positive value positively. I can see these values in the below picture of the results on the left column.



Coefficients

(Intercept)	diagnosisM	texture_mean	perimeter_mean	area_mean	smoothness_mean
0.4606835	-0.0137459	0.0006519	0.1365224	0.0007367	1.3821097
compactness_mean	radius_se	concavity_mean	concave.points_mean	symmetry_mean	fractal_dimension_mean
-3.8591198	0.1718394	-1.7216037	0.5914781	0.274664	2.0976799

Then I calculate the ME, RMSE and MAPE values to compare this model versus the models where I drop variables.

Reducing Complexity

The initial prediction of our model is based on the linear regression that is implemented on the training data to predict the evaluation. In order to decide what variables would yield a more accurate prediction, I used the drop1 function to drop columns that does not meet the benchmark (which could be seen in the <None> row) of -2250.6 in AIC. I remove the column that does not meet the benchmark and has the lowest AIC one by one until I am left with columns that meet the benchmark. After getting rid of the variables that does not meet the benchmark, I should get a better model after all of the columns that does not meet the benchmark has been dropped.

Values	**Original Model**	**1. Reduced Model**
ME	0.01130	0.01027
RMSE	0.06690	0.06643
MAPE	0.32610	0.32955

Table 1.: ME, RMSE and MAPE values for the linear regression model after each variable reduction. The removal order is: texture_mean, concave.points_worst, area_se-smoothness_se, symmetry_worst, concavity_worst, concave.points_mean, diagnosis, fractal_dimension_se, compactness_se, fractal_dimension_worst, fractal_dimension_mean, symmetry_mean.

I also calculated the ME, RMSE and MAPE values for the reduced model and I compared them to the original linear regression model to validate whether I could, indeed, improve the prediction. Then I compared the values to make sure that the test model for all the dataset in which each value represents a single regression model along with its ME, RMSE and MAPE, which appear in Table 1 above. The 13th model which excludes texture_mean, concave.points_worst, area_se-smoothness_se, symmetry_worst, concavity_worst, concave.points_mean, diagnosis, fractal_dimension_se, compactness_se, fractal_dimension_worst, fractal_dimension_mean, symmetry_mean produced the values that are slightly better than the original. Therefore, I chose the 1st model to carry the task since it is the best model.

Cross-validation

Cross-validation	Original Model	Reduced Model
ME	-0.00007908	0.0001016
RMSE	0.06459	0.06221
MAPE	0.30302	0.29613

Table 2.: Cross-validation results for the leave-one-out method on both the Original and the Reduced variable model.

I performed a Cross-Validation to make sure that the test model for all the dataset in which each value represents a single regression model along with its ME, RMSE and MAPE. Which appear in table 2 above. The 1st model which excludes texture_mean, concave.points_worst, area_se-smoothness_se, symmetry_worst, concavity_worst, concave.points_mean, diagnosis, fractal_dimension_se, compactness_se, fractal_dimension_worst, fractal_dimension_mean, symmetry_mean produced values that are slightly better than the original model (1st model) and it is the best model as well. The model created sufficiently predicts the radius_means of Breast Cancer patients in Wisconsin dataset based on a variety of factors.

Conclusion

The model created sufficiently predicts the radius_means of Breast Cancer patients in Wisconsin dataset based on a variety of factors. I came to this conclusion by looking at the root square mean error. The closer it is to zero, the better the model is at predicting, therefore, 0.06459 is sufficient.

These variables include:

- perimeter_mean: mean size of the core tumor
- area_mean
- smoothness_mean: mean of local variation in radius lengths
- compactness_mean: mean of perimeter^2 / area - 1.0
- concavity_mean: mean of severity of concave portions of the contour
- radius_se*
- texture_se*
- perimeter_se*
- concavity_se*
- concave.points_se*
- symmetry_se*
- radius_worst*
- texture_worst*
- perimeter_worst*
- area_worst*
- smoothness_worst*
- compactness_worst*

*The mean, standard error and "worst" or largest (mean of the three largest values) of these features Ire computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

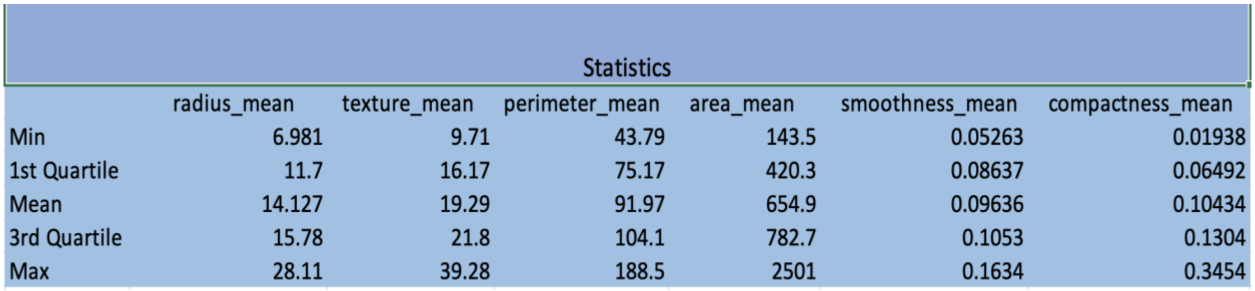
Logistic Regression

Introduction

The purpose of this method is to develop and test several logistic regression models to implement a prediction model for the likelihood of a patient in Wisconsin Breast Cancer dataset to have either a benign or malignant tumor. The columns of the data include: ID number, Diagnosis: the diagnosis of breast tissues (M = malignant, B = benign), radius_mean: mean of distances from center to points on the perimeter, texture_mean: standard deviation of gray-scale values, perimeter mean: mean size of the core tumor, area mean, smoothness mean: mean of local variation in radius lengths, compactness_mean: mean of perimeter^2 / area - 1.0, concavity mean: mean of severity of concave portions of the contour. concave points mean: mean for number of concave portions of the contour The objective of this analysis is to predict whether a person has benign or malignant tumor.

Data loading

As a first step I load the Breast Cancer Wisconsin data that was acquired from Kaggle, and I examine the summary of the dataset. The aggregations below show us the average, minimum, and maximum values in the whole dataset. This allows us to quickly see if data cleaning is needed (for example one factor has an impossibly high or low value).



Statistics

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
Min	6.981	9.71	43.79	143.5	0.05263	0.01938
1st Quartile	11.7	16.17	75.17	420.3	0.08637	0.06492
Mean	14.127	19.29	91.97	654.9	0.09636	0.10434
3rd Quartile	15.78	21.8	104.1	782.7	0.1053	0.1304
Max	28.11	39.28	188.5	2501	0.1634	0.3454

After having looked at the data, it was noticed that there are issues with the last columns named X. It has no values stored in it, therefore, it was excluded along with many columns that are not necessary for the model including the ID columns which is just an identifier. I then split the dataset into a training set (60% of the whole dataset) and the evaluation set (40%).

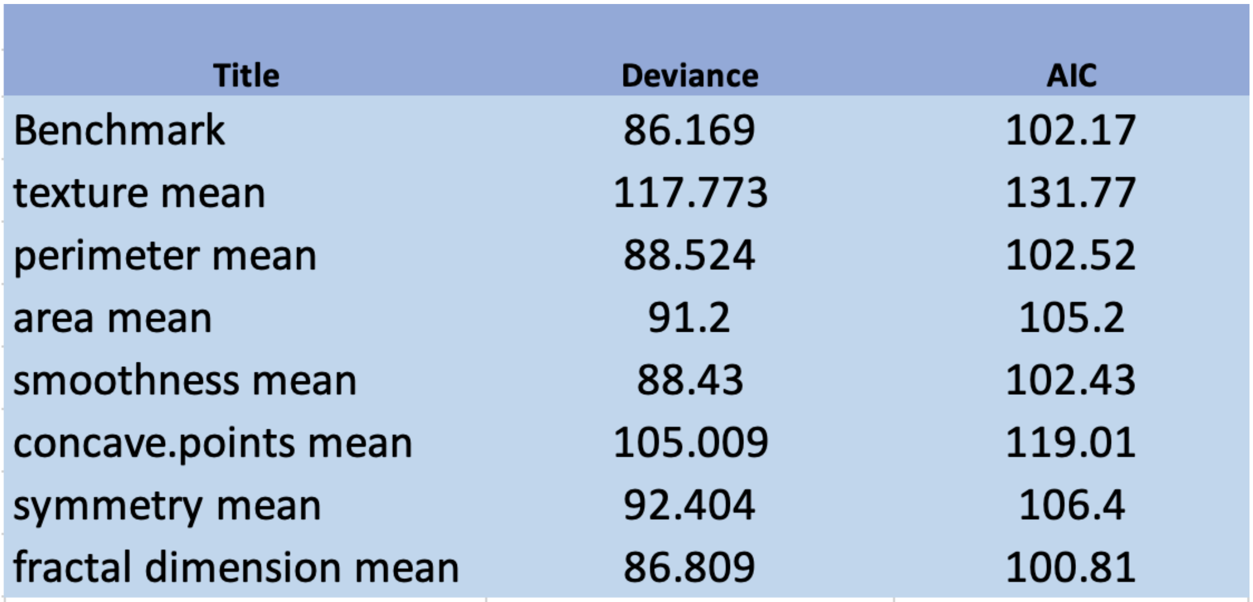
Implementation of Logistic Regression

I implemented a logistic regression model for predicting the likelihood of a patient having benign or malignant tumor. Unlike linear regression, logistic regression would predict True or False result. (no continuous value would be given)

Reducing Complexity

The initial prediction of the model is based on the logistic regression that is done on the training data to predict the evaluation. To decide what variables would provide a more accurate prediction, I used the drop1 function to drop columns that does not meet the benchmark of AIC (107.75). By looking at the values of AIC in table 1.1, I could decide which columns to be dropped first. The one that has the lowest AIC value that is also lower than the benchmark mentioned above will be dropped first. I then repeated the drop1 again to find what columns still don’t meet the benchmark of AIC and that one will be dropped. This process will be repeated until all I'm left with are columns that meet the benchmark of AIC. I eventually used drop1 and found that the rest of the columns met the benchmark. This can be seen by looking at Table 3

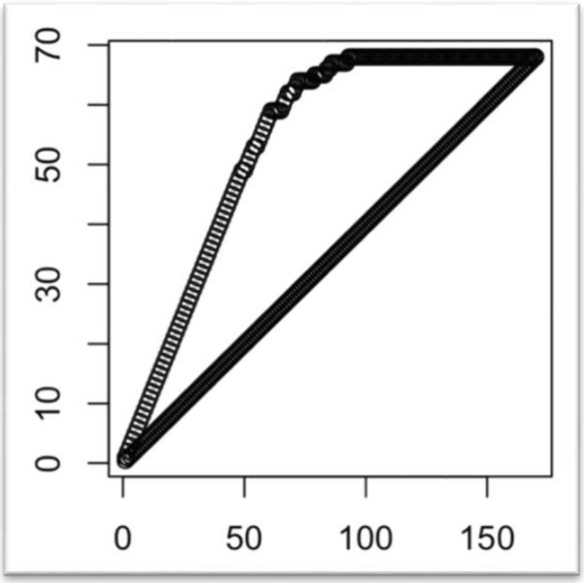
Table 3



Title	Deviance	AIC
Benchmark	86.169	102.17
texture mean	117.773	131.77
perimeter mean	88.524	102.52
area mean	91.2	105.2
smoothness mean	88.43	102.43
concave.points mean	105.009	119.01
symmetry mean	92.404	106.4
fractal dimension mean	86.809	100.81

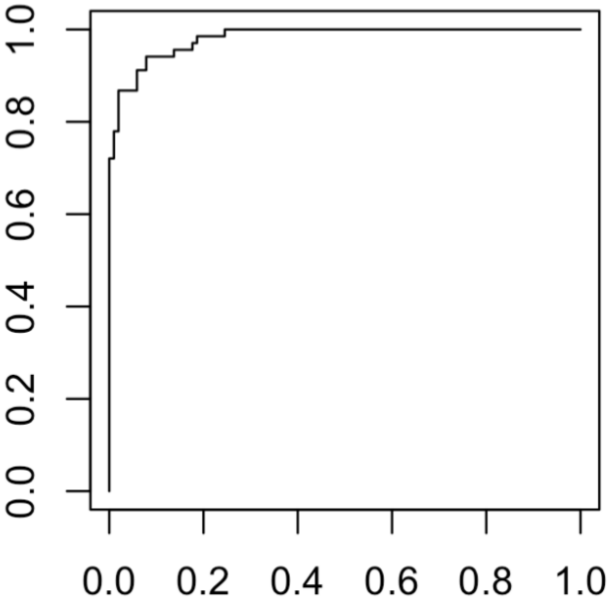
To assess how successful the model is, I calculated the errors. In the evaluation set with 170 participants, the model wrongly predicted 2 as false positives (meaning, predicted a person to have malignant tumor when they did not have it), and 9 as false negatives. The cut off value was decided to be 50%. This gives us an error rate of 6.47% and shows that the model tends to wrongfully predict a person to not have a malignant tumor when they actually do more often than the other way around. That is not good in this case, therefore, I suggest choosing a cutoff that decreases the likelihood of this.

The lift curve, which can be seen below, is a visual aid for measuring the effectiveness of the model. The x-axis shows the number of cases and the y-axis shows the number of successes.. A good lift curve should have a very steep curve at the beginning. The fact that the curve is moderately steep signifies more errors in the model.



The ROC curve plots a model's sensitivity on the vertical axis against the false positive rate on the x-axis. Unlike the lift curve, it shows percentages. The straight black line shows what results could be produced by a person randomly assigning properties to observations. Performance is better for curves that are closer to the top-left corner.

In the case of this model, the curve is significantly higher than the straight line, on top of that it is very close to the top-left corner. This signifies that the model is definitely useful, but not always accurate.



Cross Validation

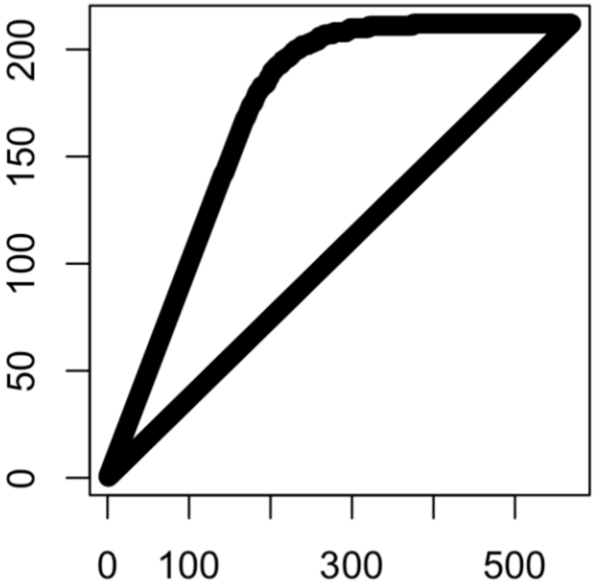
In order to check how the model performs on the whole dataset, and not just on a selected part of it, I performed leave-one-out cross-validation. The results for the model after cross- validation, for 569 observations, are as follows:

Number of false positives: 15

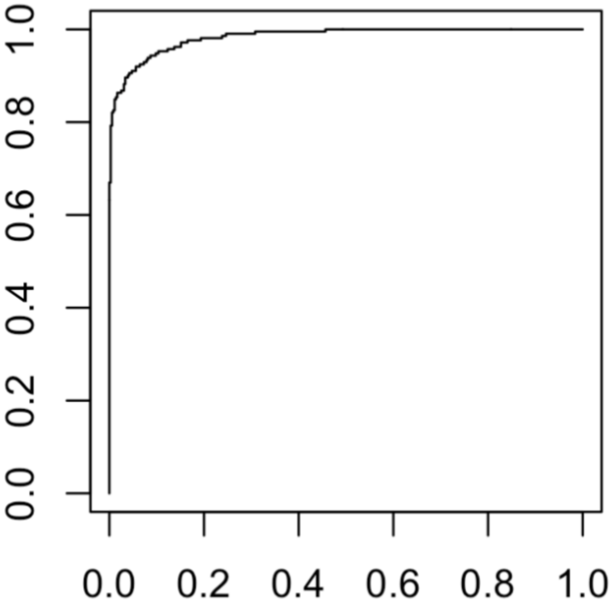
Number of false negatives: 21

Error rate: 6.32%

Lift curve



ROC curve



The higher error rate of the model after cross-validation might signify overfitting, However, the difference is very small (0.15%) which does not make a big difference so no strong evidence of overfitting. Consequently, the effectiveness of the model is as described before cross-validation.

In conclusion, the two methods used on the dataset could be of major help in aiding the decision of the need for biopsy for a breast cancer patient. However, they should be used with caution not to deny care to a patient who needs it because the system could be wrong in some cases producing false negatives when a patient is really positive. Nonetheless, this system should be used and improved using the breakthrough technology and methods that arise with time to improve its effectiveness and accuracy.

References:

Centers for Disease Control and Prevention. n.d. Basic Information About Breast Cancer. [online] Available at: https://www.cdc.gov/cancer/breast/basic_info/index.htm

Kaggle.com. 2017. Breast Cancer Wisconsin (Diagnostic) Data Set. [online] Available at: <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>

Mayo Clinic. 2019. *Breast Cancer - Symptoms And Causes*. [online] Available at: <https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470>