Andrea Carolina Flores Ramirez
A01350993
Intelligent Systems Technologies Project

# Breast Cancer Prediction with logistic regression

*Introduction*

Machine learning is one of the most important tendences of future technologies, it is used to make machines learn with different techniques, one of them is Logistic Regression which is a statistical method for predicting binary classes, it is also one of the simplest and most widely used Machine Learning algorithms for classifying two classes. For this project I created a model to predict if a woman has breast cancer or not by using different medical exam results, related to breast form and texture. As it is said, this model will require a boolean result, that is why I decided to use logistic regression for my prediction.

To understand how this prediction works you need to understand why this was an interesting topic from the medical point of view. Cancer is a central nervous system disorder that affects movement and often causes tremors. Damage to nerve cells in the brain causes a drop in dopamine levels, which causes the symptoms of Parkinson's disease. Cancer disease begins in the cells of the breast. Breast cancer can occur in women and rarely in men. Symptoms of breast cancer are lumps on the breast, blood discharge from the nipple, and changes in the shape or texture of the nipple or breast. Treatment depends on the stage of the cancer. It may consist of chemotherapy, radiation therapy, or surgery.

*Justification*

The main reason why I decided to make a prediction about this topic is because my grandma had breast cancer and she lost a breast because of a late diagnosis. Breast Cancer disease is difficult to diagnose in early stages because many women decide to go to the doctor until aggressive symptoms appear, this happens because the only home made test is to touch your breast for bumps that sometimes are confused with breast swelling for menstruation.

*Objective*

Predict if a person has Breast Cancer.

*Procedure*

For this process i downloaded a Breast Cancer database from UCI machine learning repository, this dataset use the next 34 attributes:

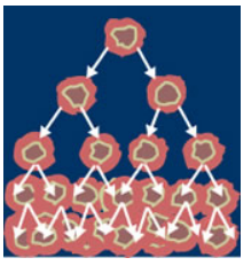
Andrea Carolina Flores Ramirez
A01350993
Intelligent Systems Technologies Project

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

# Breast Cancer Wisconsin (Prognostic) Data Set

*Download*: Data Folder, Data Set Description

**Abstract**: Prognostic Wisconsin Breast Cancer Database



| | | | | | |
|---|---|---|---|---|---|
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 198 | **Area:** | Life |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 34 | **Date Donated** | 1995-12-01 |
| **Associated Tasks:** | Classification, Regression | **Missing Values?** | Yes | **Number of Web Hits:** | 235187 |

Inside the machine repository link are different versions of the data set. The original one had 198 instances but they uploaded a new version with more instances, they also deleted two attributes that were empty so i decided to use that version.

1. ID (Patient Number)
2. Diagnosis (M = malignant, B = benign)
3. Radius mean  (mean of distances from center to points on the perimeter)
4. Texture mean (standard deviation of gray-scale values)
5. Perimeter mean
6. Area mean
7. Smoothness mean (local variation in radius lengths)
8. Compactness mean (perimeter^2 / area - 1.0)
9. Concavity mean (severity of concave portions of the contour)
10. Concave points mean (number of concave portions of the contour)
11. Symmetry mean
12. Fractal dimension mean ("coastline approximation" - 1)
13. Radius Standard error
14. Texture Standard error
15. Perimeter Standard error
16. Area Standard error
17. Smoothness  Standard error
18. Compactness  Standard error
19. Concavity Standard error
20. Concave points Standard error
21. Symmetry Standard error
22. Fractal Dimension  Standard error
23. Radius worst (mean of the three largest values)
24. Texture worst (mean of the three largest values)

Andrea Carolina Flores Ramirez
A01350993
Intelligent Systems Technologies Project

25. Perimeter worst (mean of the three largest values)
26. Area worst (mean of the three largest values)
27. Smoothness worst (mean of the three largest values)
28. Compactness worst (mean of the three largest values)
29. Concavity worst (mean of the three largest values)
30. Concave points worst (mean of the three largest values)
31. Symmetry worst (mean of the three largest values)
32. Fractal dimension worst (mean of the three largest values)

As we can see i have a couple of features that are similar, this means these features can be correlated so i need to check what are the columns that i can drop. Thanks to a heatmap I can know which are the ones that I don't need. Also, I needed to clean the cvs because the original dataset had missing values, change B to 0 (healthy) and M to 1 (cancer case) and set a column title for each attribute by hand.

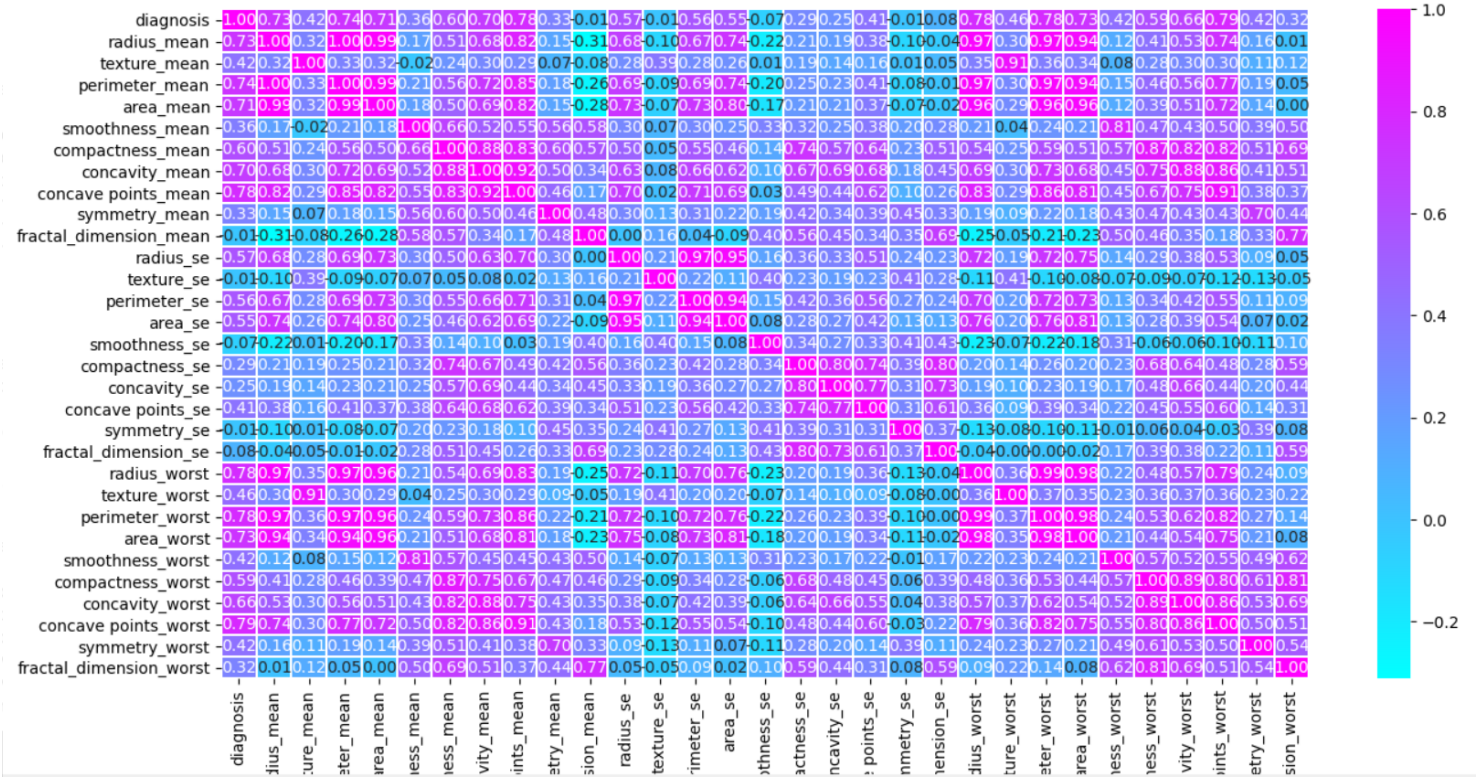



Figure 1. Heat map for Breast Cancer data set.

After this process I proceed to drop the columns that have a value of 1.0 inside the heat map just for SE and Worst, these variables represent the same principal variables but with standard error and an average of the highest 3 values. The problem with this data set is that it is really old, so its documentation is not complete, for example variables cataloged as “Worst” have a description of “average of the 3 highest values” but it doesn't say which variables so for user data test propurses i decided to drop DE and Worst attributes, before that i applied my model with all the variables, accuracy was a little bit high.

Andrea Carolina Flores Ramirez
A01350993
Intelligent Systems Technologies Project

```
Confusion matrix results for hand test model 20%
[[56 11]
 [ 5 42]]
Accuracy on test set by our model       :   86.72566371681415
```

Figure 2. Result using all variables, random state=0. 56 true positives, 11 false positives, 5 false negatives and 42 true negatives. Accuracy of 86.72%

After I decided to drop those columns I analysed what are the possible cases for all patients. I ended up with 357 Cancer cases and 212 healthy patients.

```
Counting positive and negative values:
0    357
1    212
```

Figure 3. Positive and negative values for diagnosis.

Next step was to start with the logistic regression by hand process, for this I decided to create 4 different functions inside a class named LogitRegression. This four functions were:

1. Function "__init__" is the first one, this one saves learning rate and iterations used to complete the process.
2. Function "fit" is the second function and the one in charge to complete the training, inside this function the gradient descent process with the iterations that i defined in the first function, this one uses the third function as part of the gradient descent process.
3. Function "update_weights" is the third function and the one in charge of gradient descent calculations, this one also calculates gradients and updates the original values. This one is more a helper function to complete the second one that's why I call this function inside the second function.

$$J'(\theta) = \begin{bmatrix} \frac{dJ}{dw} \\ \frac{dJ}{db} \end{bmatrix} = [\dots] = \begin{bmatrix} \frac{1}{N}\sum 2x_i(\hat{y} - y_i) \\ \frac{1}{N}\sum 2(\hat{y} - y_i) \end{bmatrix}$$

Figure 4. Gradient descent function.

4. Function "predict" is the fourth function and is the last one i se. In this function I make the logistic regression model calculation to get the final value.

$$\hat{y} = h_\theta(x) = \frac{1}{1 + e^{-wx+b}}$$

Figure 5. Logistic regression model function.

Now it's time to apply my model with the clean dataset. For this i split the train and test, initially i used 30% for test and 70% for test but i realized that a smaller test will give me a high accuracy, only if it was bigger than 15 and smaller than 25, so at the end i decided to use a test of 20% and a train of 80%. As I already said I used all attributes at the beginning with a random state of 0 and it gives me a high result (Figure 2). As I had to test with the first variables for practical purposes I applied the same conditions with the cleaned data base and I got a low accuracy so I decided to use a random state of 42 and my accuracy got higher. Stills lower than using all attributes but i don't have enough information to calculate the SE and Worst attributes because of incomplete index data. I also decided to print the confusion matrix to check in

a more detailed way my model behaviour. For this part I decided to use a learning rate of .01 and 1000 iterations as the initial value and they worked well for both cases.

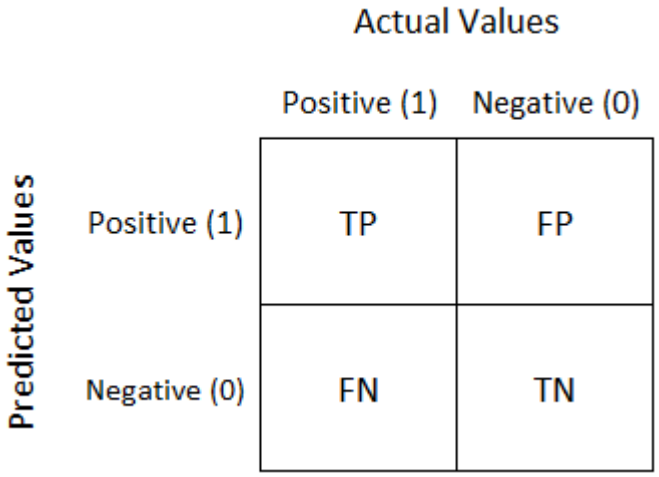



Figure 6. Confusion matrix model.

```
Confusion matrix results for hand test model 20%
[[71  0]
 [18 25]]
Accuracy on test set by our model:   84.21052631578947
```

Figure 7. Result for initial attributes with random state of 42. 71 positive positives, 0 false positives, 18 false negatives and 25 true negatives. Accuracy of 84.21%.

Now it's time to test my model with user data and calculate compactness attribute as the data set description says.

*Results*

```
Answer the next questions:

Radius or mean of distances from center to points on the perimeter: 12.89
Texture or standard deviation of gray-scale values: 13.12
Perimeter: 81.89
Area: 515.9
Smoothness or local variation in radius lengths: .06955
Severity of concave portions of the contour: .0226
Number of concave portions of the contour: .01171
Symmetry mean: .1931
Fractal dimension mean: .5796
Results will be printed with 1 as Breast Cancer case and 0 as healthy
Hand Prediction:
[0]
```

Figure 8. Healthy input data prediction. Result: True negative.

Andrea Carolina Flores Ramirez
A01350993
Intelligent Systems Technologies Project

```
Answer the next questions:

Radius or mean of distances from center to points on the perimeter: 28.11
Texture or standard deviation of gray-scale values: 39.28
Perimeter: 188.5
Area: 2501
Smoothness or local variation in radius lengths: .1447
Severity of concave portions of the contour: .4268
Number of concave portions of the contour: .2012
Symmetry mean: .304
Fractal dimension mean: .09744
Results will be printed with 1 as Breast Cancer case and 0 as healthy
Hand Prediction:
[1]
```

Figure 9. Cancer case input data prediction. Result: True positive.

```
Answer the next questions:

Radius or mean of distances from center to points on the perimeter: 15.46
Texture or standard deviation of gray-scale values: 19.48
Perimeter: 101.7
Area: 748.9
Smoothness or local variation in radius lengths: .1092
Severity of concave portions of the contour: .1466
Number of concave portions of the contour: .08087
Symmetry mean: .1931
Fractal dimension mean: .05796
Results will be printed with 1 as Breast Cancer case and 0 as healthy
Hand Prediction:
[0]
```

Figure 10. Healthy input data prediction. Result: True negative.

Bibliography:


- Dr. William H. Wolberg, General Surgery Dept. University of Wisconsin, Clinical Sciences CenterMadison, WI 53792 (2021). UCI Machine Learning Repository: Breast Data Set. Retrieved May 12, 2021, from http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29