MATH 123 Final Project
Breast Cancer Analysis

Erick Giron
Major: Mathematics - Statistics option

May 17, 2022

Contents

1 Introduction

In this project, I plan on using the breast cancer data set from UCI machine learning repository for my analysis. The questions I plan on answering are of the following:

- Are people going through menopause more susceptible to receive breast cancer then those who aren't?
- Are younger women more susceptible to get breast cancer than older women?
- Can we predict the breast cancer reoccuring based on if the patients are going through menopause or if they have gone through radiation therapy?

1.1 Properties

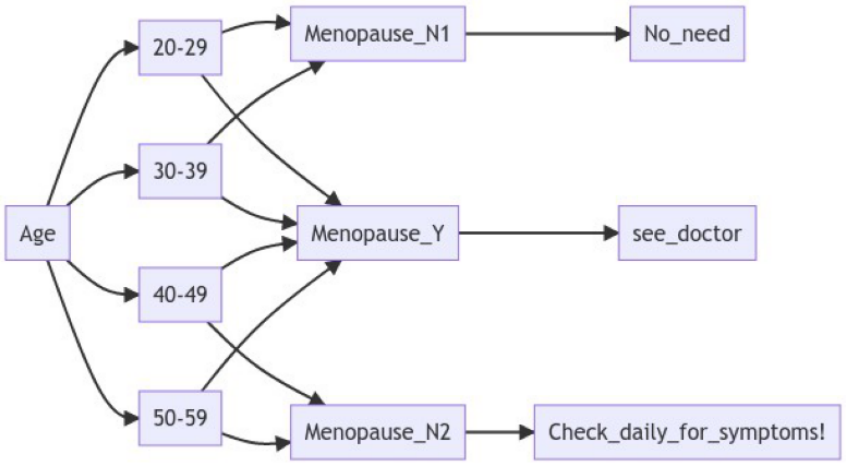
- Class: no-recurrence-events, recurrence-events
- Age: age of the patient at the time of diagnosis
- Menopause: whether the patient is pre- or postmenopausal at time of diagnosis
- Tumor-size: the greatest diameter (mm) of the excised tumor
- Inv-nodes: the number of auxiliary lymph nodes that contain metastatic breast cancer visible on histological examination
- Node-caps: if the cancer does metastise to a lymph node, although outside the original site of the tumor it may remain "contained" by the capsule of the lymph node. However, over time, and with more aggressive disease, the tumor may replace the lymph node and then penetrate the capsule, allowing it to invade the surrounding tissues
- Degree of malignancy: the historical grade (range 1-3) of the tumor. Tumors that are grade 1 predominantly consist of cells that , while neoplastic, retain many of their usual characteristics. Grade 3 tumors predominately consist of cells that are highly abnormal
- Breast: the breast in which the cancer occurs
- Breast-quad: the quadrants of the breast where the nipple is used as the central point
- Irradiation: radiation therapy is a treatment that uses high-energy x-rays to destroy cancer cells.

2 Methods and Results

In order to answer my questions, I will be using decision theory to help determine the different outcomes for each age group of whether or not they should get checked for breast cancer. To solve my more relational based questions, I will use logistic regression to create a model that could help predict future outcomes.

2.1 Decision Theory

In order to determine if a person who's age is from $20-59$ should go see a doctor based on if they experienced menopause.



Note: Depending on the age of the patient, this decision tree helps determine if the patient needs to go see a doctor if they are going through menopause to make sure that they

By looking at the results, we see that those who experience menopause, whether early or late, should go see a doctor since they have a higher chance of obtaining breast cancer. It is rare for someone between the age $20-39$ to go through menopause which is why for those who do should go see a doctor immediately to understand why they are experiencing it. For those whose age is in the range $40-55$ that haven't experience menopause should check daily for symptoms. The following is the percentage at which a women pre-menopause has to attain breast cancer

Age Group	Probability
$20-29$	.00058%
$30-39$	0.49%
$40-49$	1.55%
$50-59$	2.40%
$60-69$	3.54%

2.2 Logistic Regression

In this approach, I decided to use logistic regression to answer my more complex questions by creating a model to see if it can predict various questions.

2.2.1 Recurrence event vs Menopause

By creating a model that trained and tested my data set, we were given that $\beta_0 \approx 0.0296$ and $\beta_1 \approx 0.1106$ thus our logistic model looks like,

$$Y \approx \frac{e^{0.0296+X*0.1106}}{1+e^{0.0296+X*0.1106}} \tag{1}$$

with a given accuracy score of 0.547. The confusion matrix below helps understand our accuracy score since we see that the model only correctly predicted the proportion of women who don't have menopause will not have the cancer occur again in their lives.

	Positive	Negative
Positive	0	34
Negative	0	47

Thus by creating a ROC curve, we could determine the percentage of true positive predicted by the model.

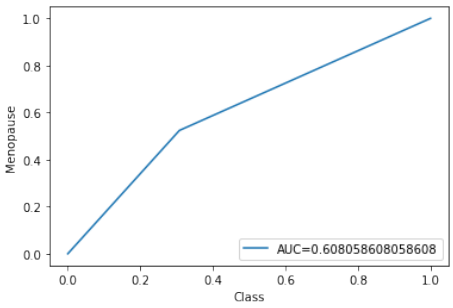


Figure 1: The AUC (area under the curve) represents the accuracy which our model has predicted our values.

2.2.2 Recurrence event vs Irradiate

In this case, I decided to create a model that could help predict the chance of recurrence class depending if the patient has gone through radiation treatment. The values that our model produced are $\beta_0 \approx -1.3871$ and $\beta_1 \approx 0.793$ which we can then form the equation,

$$Y \approx \frac{e^{-1.3871+X*0.793}}{1+e^{-1.3871+X*0.793}} \tag{2}$$

with an accuracy score of 0.453. The confusion matrix show us the proportion the model predicted correctly. This mean that those who go through radiation therapy will have the cancer reoccurring again.

	Positive	Negative
Positive	65	0
Negative	21	0

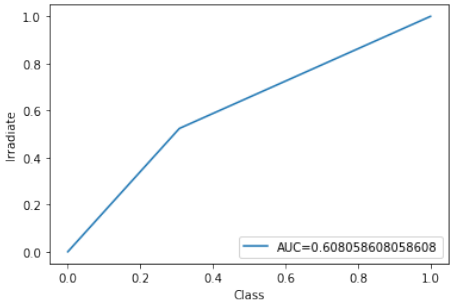


Figure 2: The AUC (area under the curve) represents the accuracy which our model has predicted our values.

2.3 Optimization

In order to optimize my model, I decided to do difference cases where I changed my test size and another where I changed my random state of the model.

2.3.1 recurrence vs menopause: test size = .4

By changing the test size to 40 percent and our train size to 60 percent, our values that were produced in our model are $\beta_0 \approx 0.1211$ and $\beta_1 \approx -0.0703$.

$$Y \approx \frac{e^{0.1211+X*-0.0703}}{1+e^{0.1211+X*-0.0703}} \tag{3}$$

	Positive	Negative
Positive	0	55
Negative	0	60

Our accuracy score for this model is 0.521.

2.3.2 recurrence vs menopause: random state

By changing the random state in our model, it chooses new random values to be in our test data set and our train data set. By doing this, my model was able to produce new coefficients $\beta_0 \approx -0.0208$ and $\beta_1 \approx 0.263$

$$Y \approx \frac{e^{-0.0208+X*0.263}}{1+e^{-0.0208+X*0.263}} \tag{4}$$

	Positive	Negative
Positive	29	10
Negative	34	13

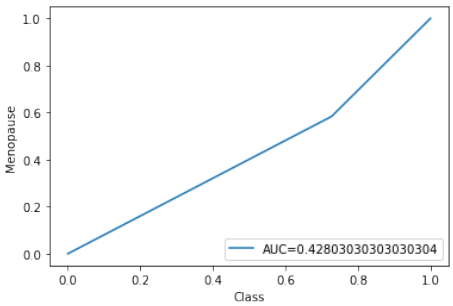


Figure 3: The AUC (area under the curve) represents the accuracy which our model has predicted our values.

Our accuracy score for this model is 0.488.

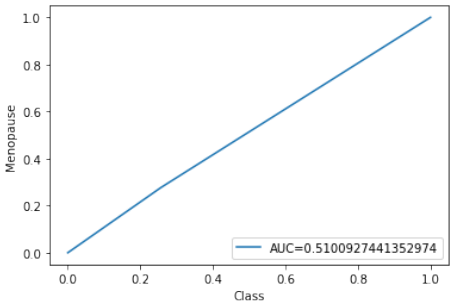


Figure 4: The AUC (area under the curve) represents the accuracy which our model has predicted our values.

2.3.3 recurrence vs irradiate: test size = .2

By changing the test size to 40 percent and our train size to 60 percent, our values that were produced in our model are $\beta_0 \approx -1.3328$ and $\beta_1 \approx 0.6843$

$$Y \approx \frac{e^{-1.3328+X*0.683}}{1+e^{-1.3328+X*0.683}} \tag{5}$$

	Positive	Negative
Positive	45	0
Negative	13	0

Our accuracy score for this model is 0.7759.

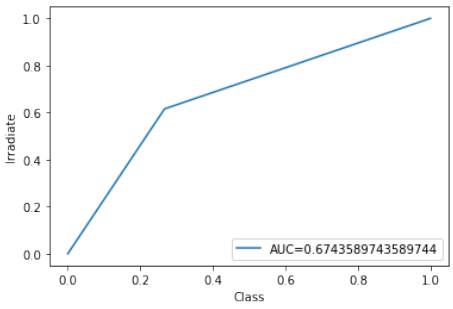


Figure 5: The AUC (area under the curve) represents the accuracy which our model has predicted our values.

2.3.4 recurrence vs irradiate: random state

By changing the random state in our model, it chooses new random values to be in our test data set and our train data set. By doing this, my model was able to produce new coefficients $\beta_0 \approx -1.4063$ and $\beta_1 \approx 0.7548$

$$Y \approx \frac{e^{-1.4063+X*0.7548}}{1+e^{-1.4063+X*0.7548}} \tag{6}$$

	Positive	Negative
Positive	65	0
Negative	21	0

Our accuracy score for this model is 0.756.

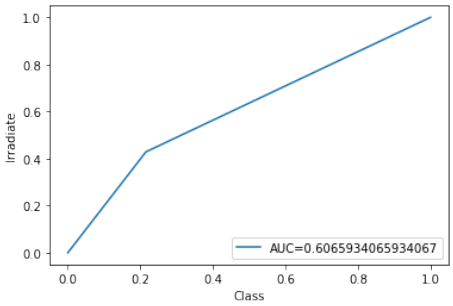


Figure 6: The AUC (area under the curve) represents the accuracy which our model has predicted our values.

3 Discussion

In this project, I wanted to help minimize the risk of breast cancer in women since they have a higher chance to obtain it as they get older. By using the

breast cancer data set from the UCI machine learning repository, I was able to create a decision tree to help women determine if they should go get checked for breast cancer depending if they are going through menopause. I then wanted to see if there was a relationship between the cancer reoccurring based on if the patients have gone through radiation therapy or menopause. I created a model to see if we could predict for any future patients. We can see when optimizing our models, it turns out that for the case with irradiation, when we changed the test sized, we got a better accuracy score, but it's probably due to the fact that our model is testing less values and training more to get a better prediction. On the other hand, it seems like optimizing our model for menopause didn't produce a better accuracy score. In this case, I plan to use different classification methods to see which model is better suited such as KNN models.

4 Reflection

I came into this class assuming that we'll be learning machine learning algorithms based on the name of the class but instead it introduced how mathematical models could be used in other areas such as economics and in habitat science such as predicting species population. This class has reinforced my motivation in learning more about how to use this basic knowledge of mathematical models to possible help a business grow or even go into research. It was really great to have you again as a teacher but also as a mentor since you sent me opportunites for programs that I would have never knew of.

References

[1] Galarnyk, M. (2022, April 27). Understanding train test split (Scikit-Learn + Python). Medium.

[2] Surakasula, A., Nagarjunapu, G. C., Raghavaiah, K. V. (2014, January). A comparative study of pre- and post-menopausal breast cancer: Risk factors, presentation, characteristics and management. Journal of research in pharmacy practice.

[3] Chan, C. (2020, December 9). What is a ROC curve and how to interpret it. Displayr.

[4] James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). An introduction to statistical learning: With applications in R. Springer.