Breast Cancer Survival Analysis

Does Age at Diagnosis Have an Effect on Breast Cancer Survival?

Christien Bryan Reyes Viray

October 8, 2024

**Table of Contents**

***Choose a Dataset ................................................................................................................................................... 3***

***Ask a Question ....................................................................................................................................................... 3***

***Graphs the Data .................................................................................................................................................... 3***

***Train and Test Algorithm ...................................................................................................................................... 6***

***Evaluate the Model ................................................................................................................................................ 7***

***Answering the Question and Conclusion ............................................................................................................. 7***

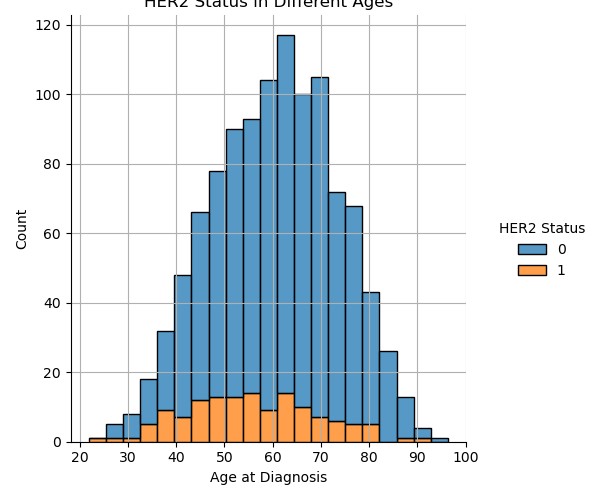
# Choose a Dataset

The dataset that has been chosen for this project is the Breast Cancer Survival Analysis dataset. This dataset contains detailed information of patients with breast cancer including their survival status. The dataset was found here: <https://www.kaggle.com/datasets/gunesevitan/breast-cancer-metabric>

# Ask a Question

Does Age at Diagnosis Have an Effect on Breast Cancer Survival?

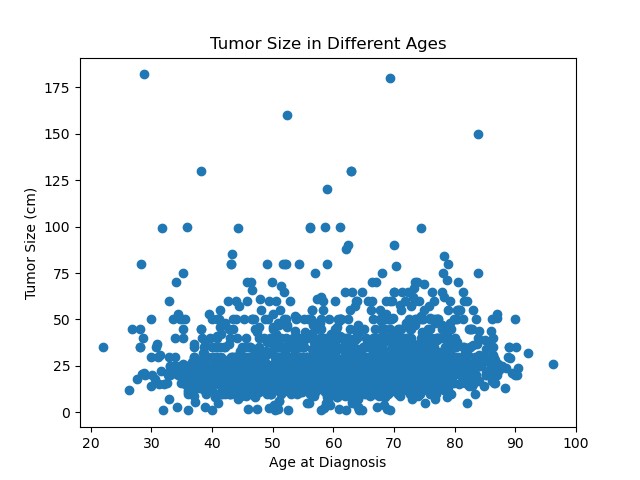
Graphs the Data



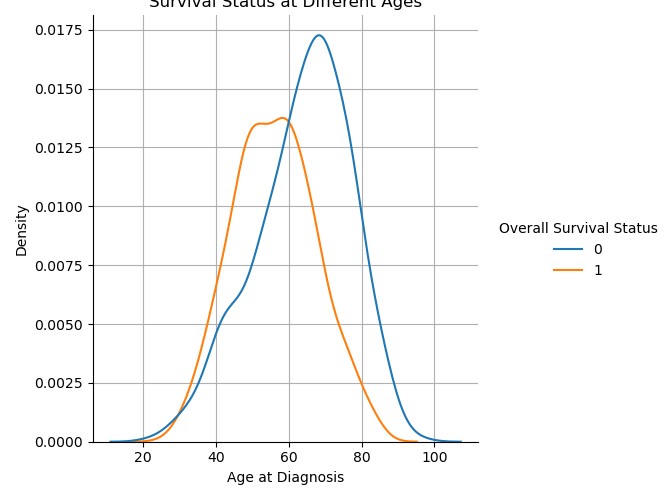
This graph looks at the relation between age at diagnosis and HER2 status. HER2 status checks the patients of their HER2 levels to see if it's higher or lower than normal. If HER2 levels are high, the status is positive (1) and if they are low the status is negative (0). HER2 positive patients have cancer that tends to grow and spread faster. This graph shows how the ages of different patients are HER2 positive, and it seems that patients around

40-65 are HER2 positive, meaning their cancer is likely to grow and spread faster than HER2 negative patients.

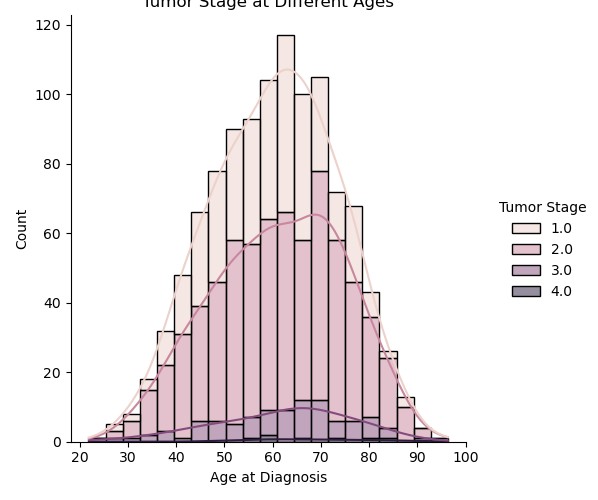
The creation of this specific graph was not discussed in class. It was found on the seaborn website from this link: https://seaborn.pydata.org/tutorial/function\_overview.html



This graph looks at the relation between age at diagnosis and tumor sizes in centimeters. This graph indicates that patients from 20-90+ all have around the same tumor sizes ranging from around 5 centimeters to 50 centimeters.

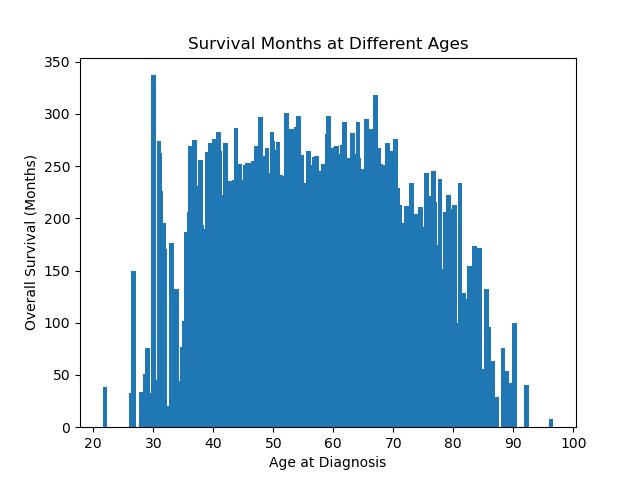


This graph looks at the relation between age at diagnosis and survival status, 0 being deceased and 1 being living. A trend line is used for both 1 and 0 to describe the trend in age and survival status. The graph indicates that from ages 30-55 the patients are more likely to survive from breast cancer. However, from ages 56-70 patients are less likely to survive breast cancer. The creation of this specific graph was not discussed in class. It was found on the seaborn website from this link: https://seaborn.pydata.org/tutorial/function\_overview.html



This graph shows the relation between age at diagnosis and tumor stages. The graph indicates that most of all patients have either stage 1 or stage 2 tumors and very few patients have stage 3 and stage 4. We can relate this graph to age of diagnosis vs tumor size. Since most patients have similar tumor sizes, we can conclude that the stage of those tumor sizes were majority stage 1 and stage 2. The creation of this specific graph was not discussed in class. It was found on the seaborn website from this link:

https://seaborn.pydata.org/tutorial/function\_overview.html



This graph shows the relation between age at diagnosis and the overall survival in months. The graph suggests that from ages 35 to approximately 70 have survived longer with breast cancer, and ages 30-32, having the longest overall survival in months. Ages from 30 and under survived the least long months, along with ages 80 and over.

# Train and Test Algorithm

|  |
| --- |
| x=bCancer.drop('Overall Survival Status',axis=1) y=bCancer['Overall Survival Status'].values  from sklearn.model\_selection import train\_test\_split x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.3,random\_state=87)  from sklearn.linear\_model import LogisticRegression logmodel=LogisticRegression() logmodel.fit(x\_train,y\_train) predictions=logmodel.predict(x\_test)  logmodel=LogisticRegression(max\_iter=2000) logmodel.fit(x\_train,y\_train) predictions=logmodel.predict(x\_test) |

This is the code I used to train and test my algorithm. First, I created my independent and dependent variables.

For x I remove the Overall Survival Status because that is what I will be using as my dependent variable (y).

Then I split my data into training and testing. I import a Logistic Regression model so that I can train it. I use the fit method so I can use the x and y training data to train the model on the dataset. Then I use the predict method to make predictions using the x variable testing data.

# Evaluate the Model

Classification Report:

precision recall f1-score support

0 0.81 0.85 0.83 175 1 0.82 0.78 0.80 153

accuracy 0.82 328 macro avg 0.82 0.81 0.82 328 weighted avg 0.82 0.82 0.82 328

The classification report concluded that the accuracy of the model for positive predictions was 82% accurate and negative predictions were 81% accurate. The recall indicated that negative cases that were correctly identified was 85% while positive cases were 78%. And the f1 score shows 83% and 80% accurate when negative and positive predictions were correct. When creating a model, the desired accuracy for the model should be 70%-90% and my model satisfies this range. One way of improving the accuracy of the model is doing more research on the columns that were dropped. By keeping some columns that could relate to others could help the model identify more trends and patterns to create a more accurate model. Another way to increase accuracy of the model is using a decision tree classifier instead of a logistic model. Using a decision tree classifier and training it with my data could increase the accuracy to as high as 100%.

# Answering the Question and Conclusion

After creating test patients to answer my question, “Does Age Have an Effect on Breast Cancer Survival?”, I have concluded that yes, age at diagnosis does influence breast cancer survival. I created test patients and inputted similar data, except for age at diagnosis. I tested the model using patients from ages 20, 30, 40, 50 and 80. I have concluded that patients under 80 have a better chance at surviving breast cancer than patients over 80. This can be due to several factors such as stronger immune systems for patients under 80.