

A grayscale image of a hand holding a bright pink ribbon, which is a symbol for breast cancer awareness. The background is a dark, textured wooden surface. The title text is overlaid on the right side of the image.

Breast Cancer Survival Prediction

Nakisa Abbasi



Introduction

❖ Why and Who Cares?

- **Breast cancer is among the most common cancers and a common cause of death among women**
- **The Patients want to know is how long they will survive**
- **Patients with different criteria can have very different survival rates.**
- **We will predict survival rate for patients with different diagnostic criteria.**



Data Information

- Data is acquired from Seer Breast Cancer Data
- Was obtained from the 2017 November update of the SEER, from cancer patients
- From female patients with infiltrating duct and lobular carcinoma breast cancer diagnosed in 2006-2010.



Data Information

Each record is consist of:

- ❖ **age, race, marital status, t stage, n stage,a stage, tumor size, estrogen status, progesterone status,regional nodes examined, regional nodes positive and status of the patient.**



Data wrangling

Data cleaning?

This dataset is not so dirty, no Null records. Some text cleaning

- Moderately differentiated; Grade II: Grade II
- Poorly differentiated; Grade III: → Grade III
- ... →



Data wrangling

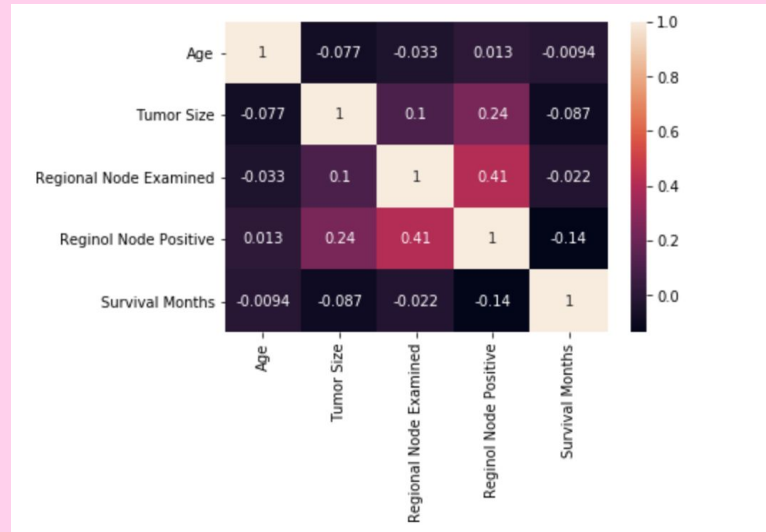
Categorical data to numeric?

- **For features “one hot encoding”**
 - `features = pd.get_dummies(features, drop_first = True)`
- **For target “LabelEncoder”**
 - `target = pd.DataFrame(le.fit_transform(target), columns = ['Status'])`



Data wrangling

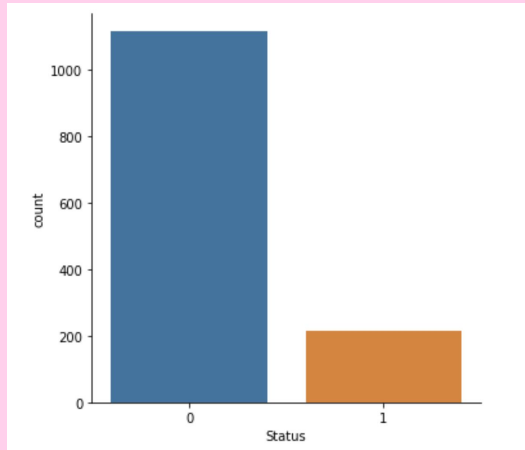
Any correlation?



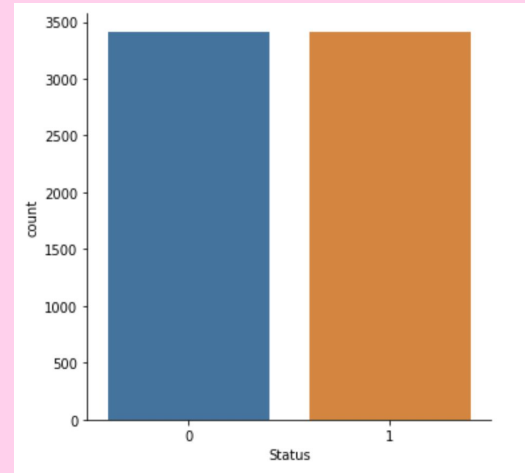


Data wrangling

Is the target data balanced?



Apply SMOTH



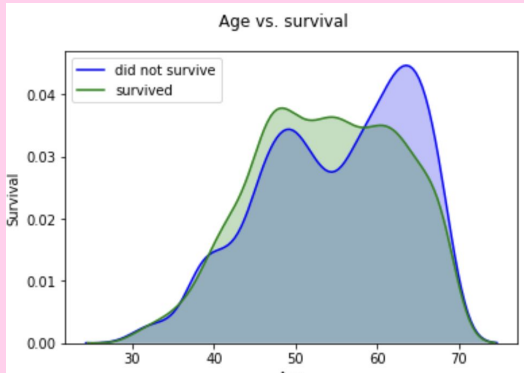


EDA

Continuous data:

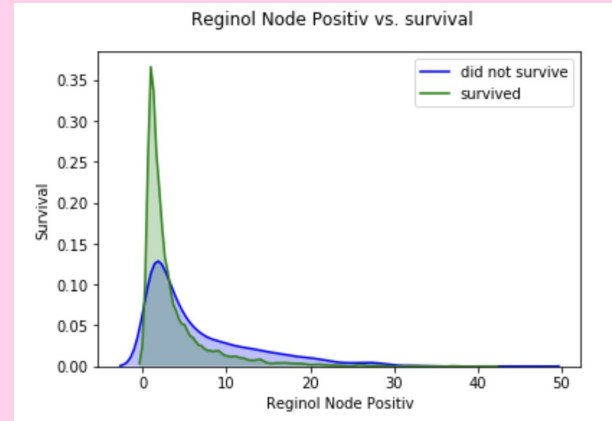
1. Age VS Survival

There is an increase in death numbers by age increase



2. Regional Node Positive VS Survival

Most of the survived patients have less number of positive nodes



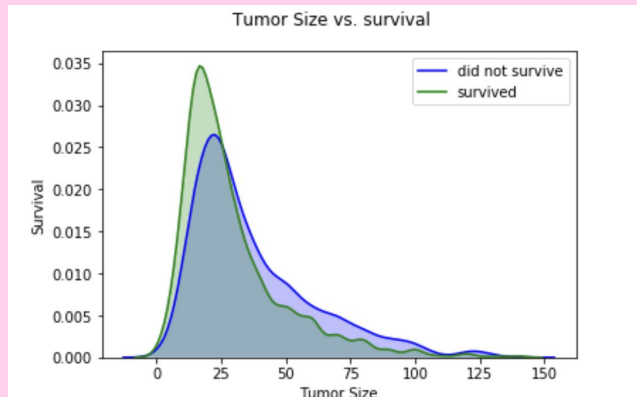


EDA

Continuous data continued:

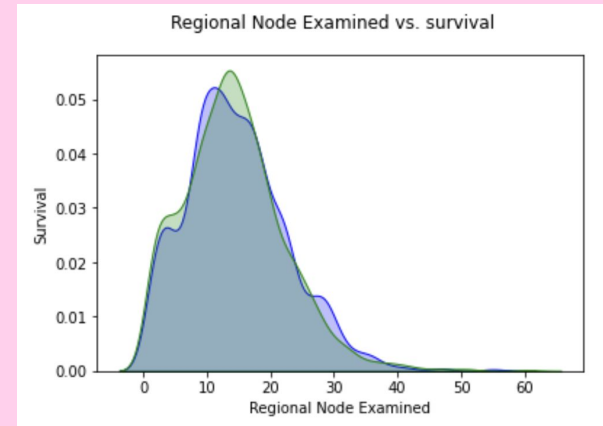
3. Tumor Size VS Survival

Most of the survived patients have smaller tumor size



4. Regional Node Examined VS Survival

The number of nodes being examined seems to be equal in most of the patient



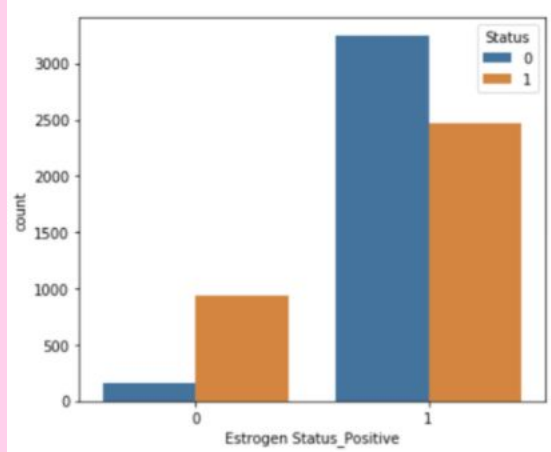


EDA

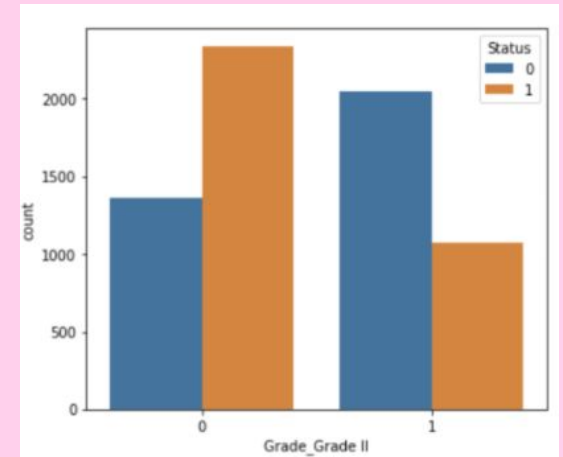
Categorical data continued:

Most important ones:

Apparently the number of survived patients is higher with Estrogen positive



The survival rate is higher with Grade 2

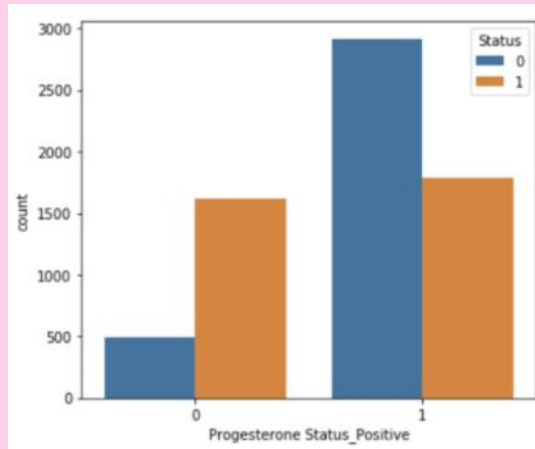




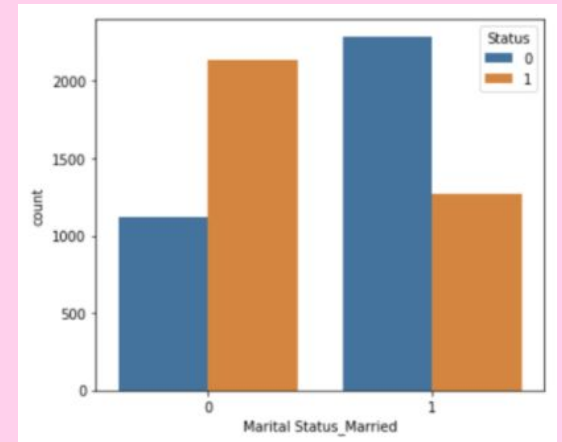
EDA

Categorical data continued:

The number of survived patients is higher with progesterone positive



Married patients have higher chance to survive





Preprocessing

- **Which Features are more important?**

Chi-squared function was used with SelectKBest from Sklearn to calculate the highest scores:

1. Survival Months
2. Tumor Size
3. Regional Node Positive
4. Grade_Grade II
5. Progesterone Status_Positive
6. 6th Stage_IIC
7. Marital Status_Married
8. Marital Status_Single
9. Race_Other
10. Estrogen Status_Positive

	Name	Score
0	Survival Months	17478.803015
1	Tumor Size	1967.014049
2	Reginol Node Positive	1750.902463
3	Grade_Grade II	212.033694
4	Marital Status_Married	182.786460
5	Progesterone Status_Positive	173.888837
6	6th Stage_IIB	118.987770
7	Marital Status_Single	106.100394
8	N Stage_N3	105.947388
9	6th Stage_IIC	105.947388
10	Race_Other	98.157095
11	Estrogen Status_Positive	64.603892
12	Age	40.265232
13	Marital Status_Widowed	37.917861
14	6th Stage_IIIA	31.855896
15	Race_White	22.984368
16	N Stage_N2	10.651684
17	A Stage_Regional	5.579573
18	T Stage_T3	5.100155
19	Marital Status_Separated	4.193361
20	Regional Node Examined	3.945062
21	6th Stage_IIB	1.389694
22	T Stage_T4	1.146433
23	Grade_anaplastic	0.243042
24	T Stage_T2	0.155767
25	Grade_Grade III	0.131553



Preprocessing

- Are the data in the same range?

Feature	Min	Max
Age	30	69
Tumor size (mm)	1	140
Regional node examined	1	61
Regional node positive f	1	46
Survival Months	1	107



Modeling

There are 5 models that we will investigate

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest
4. KNeighborsClassifier
5. Support Vector Machine



Modeling

With default values

With all available data

With the selected features