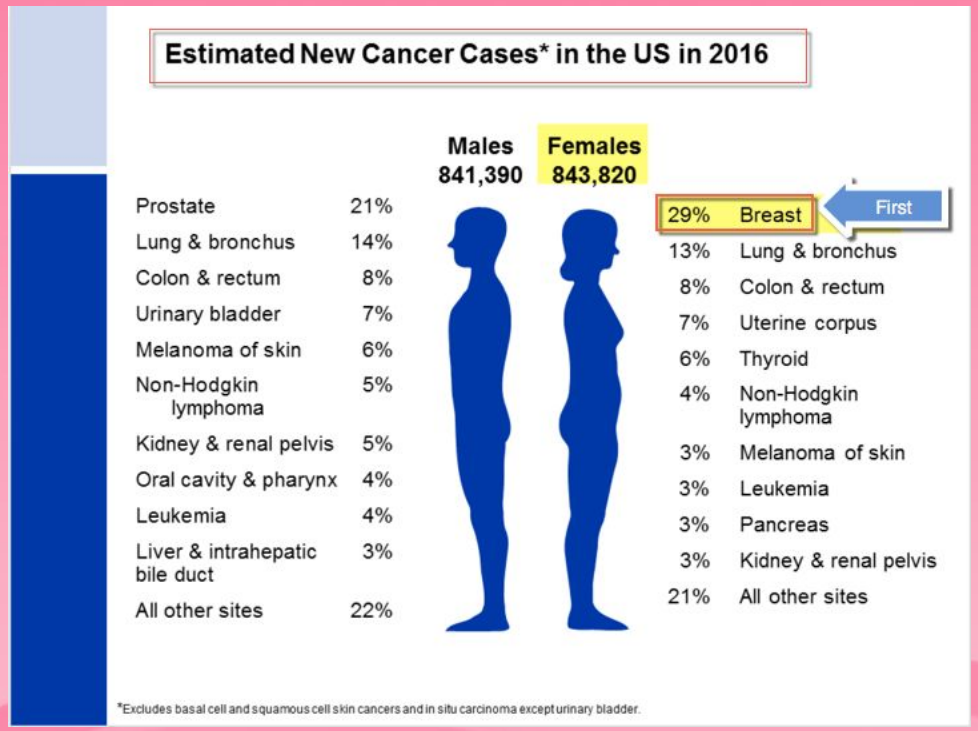
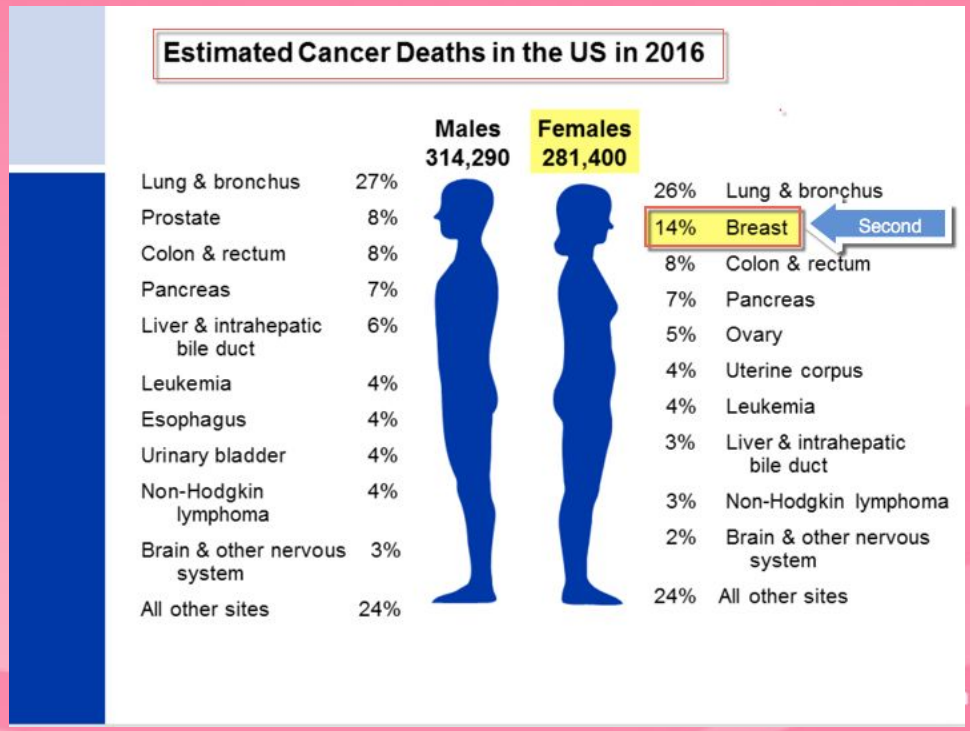
Factors that Predicts Risk of Breast Cancer in Women After a Regular Mammogram

Maria Pichardo
Capstone Project DSI GA

Overview

As per the American Cancer Society Breast Cancer is the the most common cancer among Women and the the number two cause of cancer death.



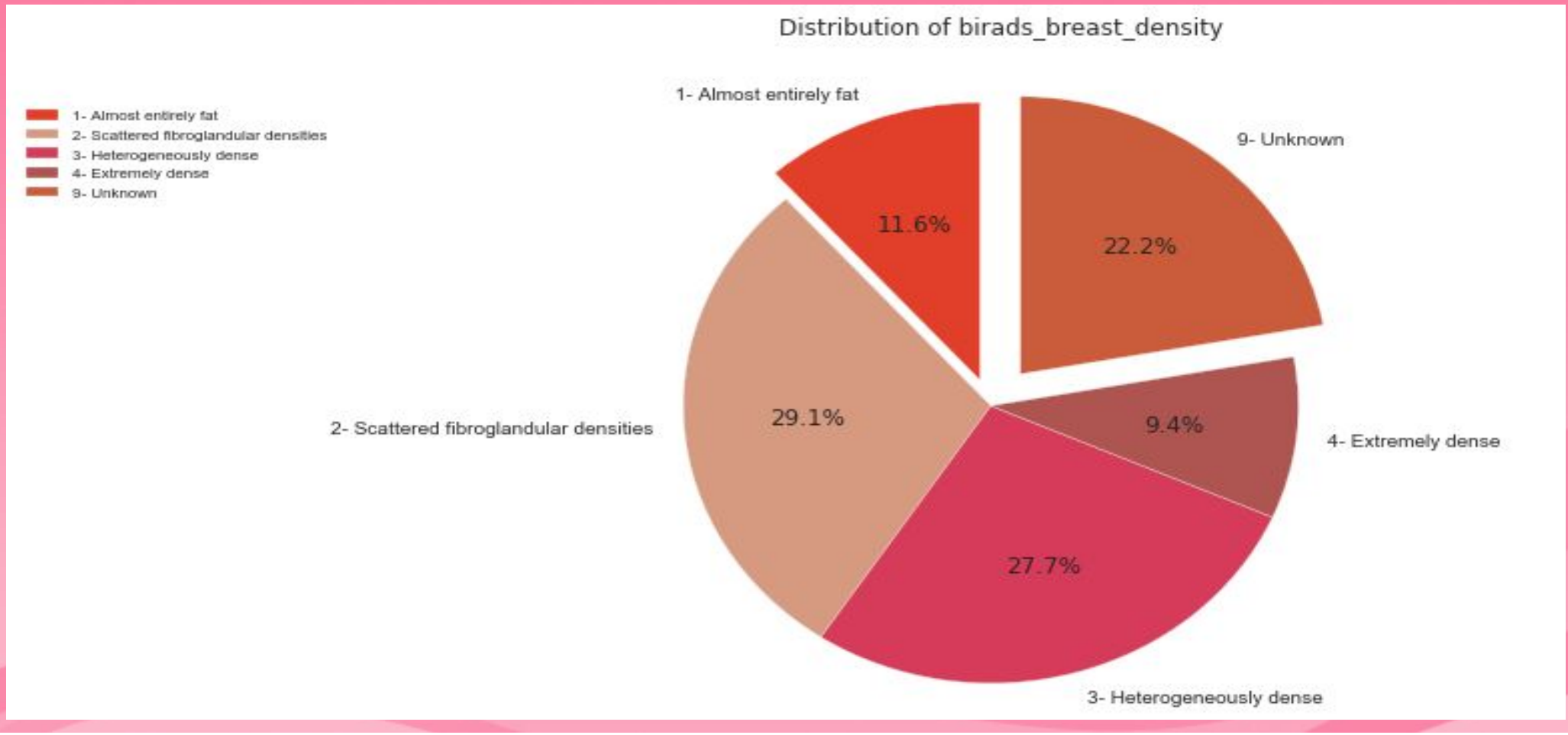


Main Objective

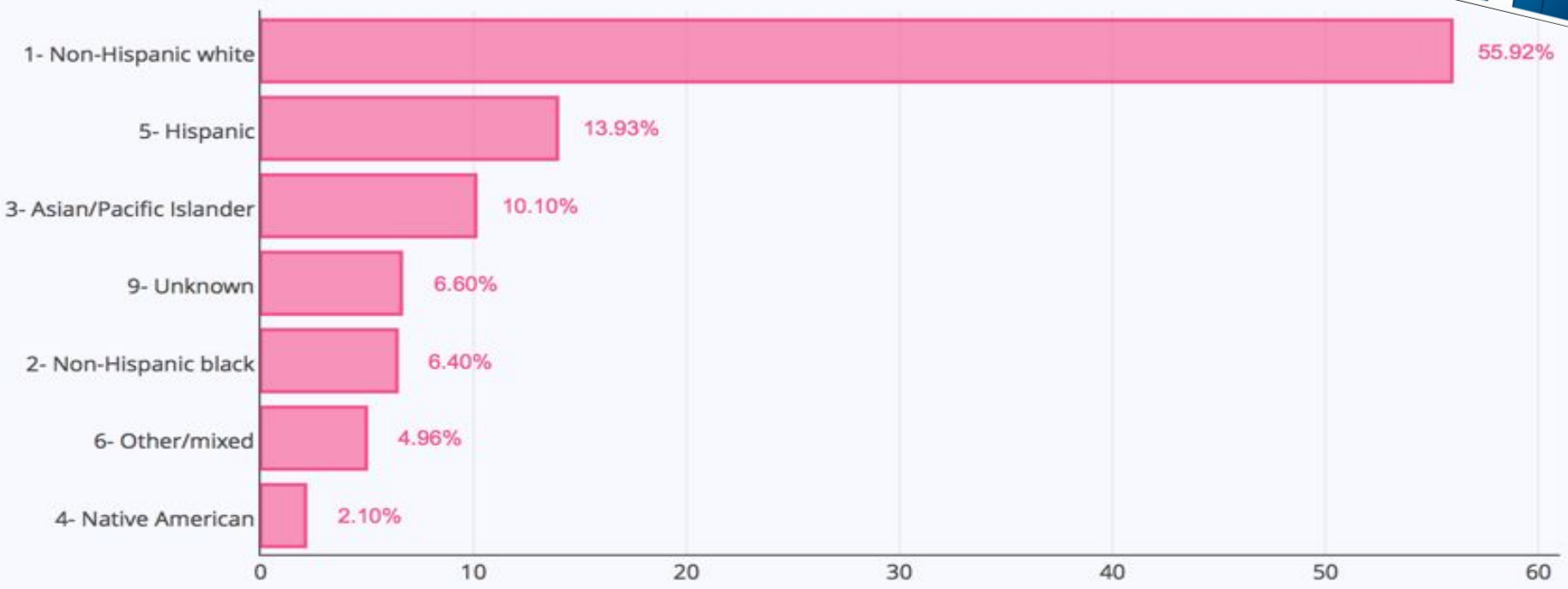
The aim of this capstone project is to determine which factors have more incidence in the breast cancer diagnosis after a regular mammogram. I used a dataset provided by the Breast Cancer Surveillance Consortium (BCSC).

Year	Calendar year of observation
age_group_5_years	Age (years) in 5 year groups
race_eth	Race/ethnicity
first_degree_hx	History of breast cancer in a first degree relative
age_menarche	Age (years) at menarche
age_first_birth	Age (years) at first birth
BIRADS_breast_density	← **Target**
current_hrt	Use of hormone replacement therapy
menopaus	Menopausal status
bmi_group	Body mass index
biophx	Previous breast biopsy or aspiration
breast_cancer_history	Prior breast cancer diagnosis
count	Frequency count of this combination of covariates

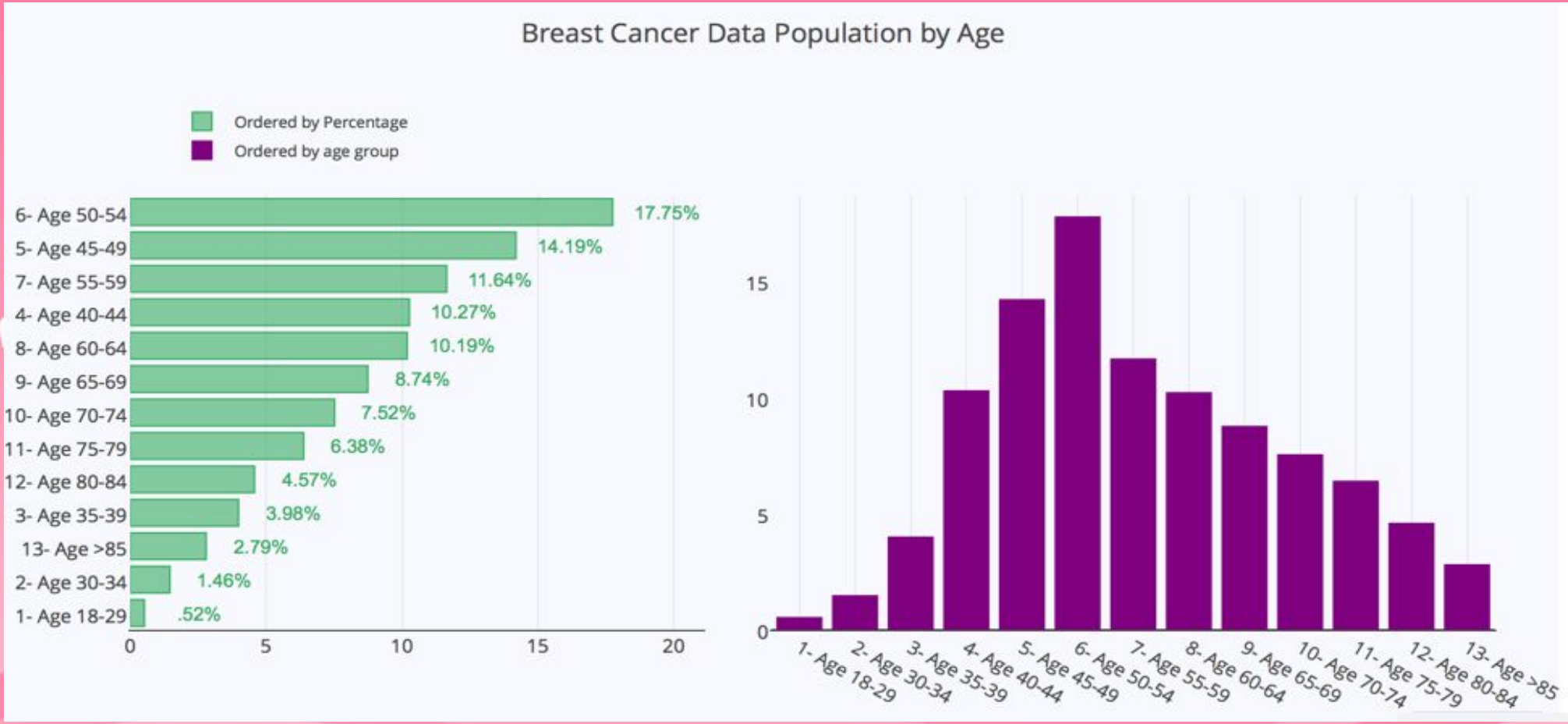
EDA - BRADS

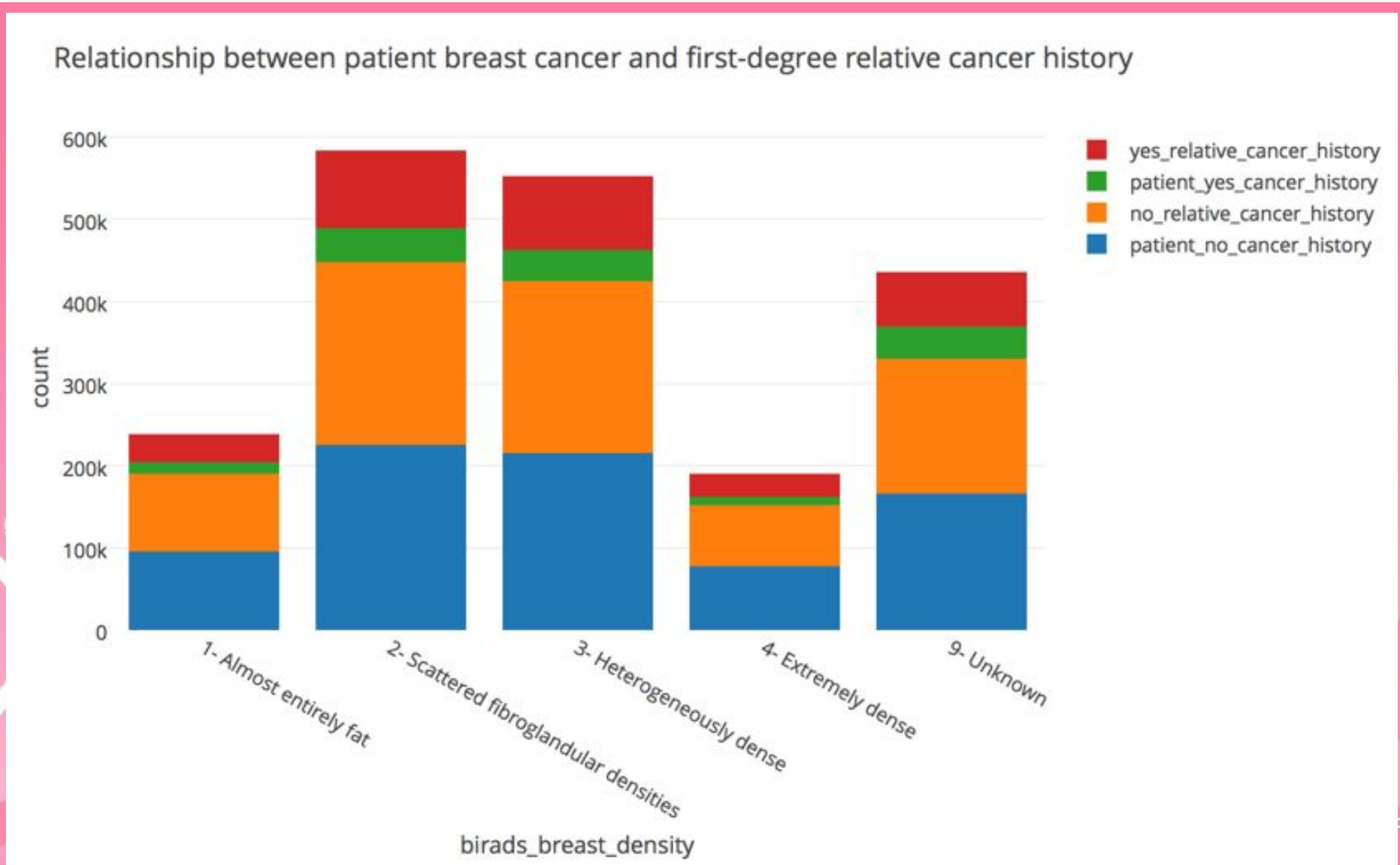

Distribution of birads_breast_density
1- Almost entirely fat
2- Scattered fibroglandular densities
3- Heterogeneously dense
4- Extremely dense
9- Unknown
11.6%
22.2%
29.1%
9.4%
27.7%
1- Almost entirely fat
9- Unknown
2- Scattered fibroglandular densities
4- Extremely dense
3- Heterogeneously dense

Race Factor

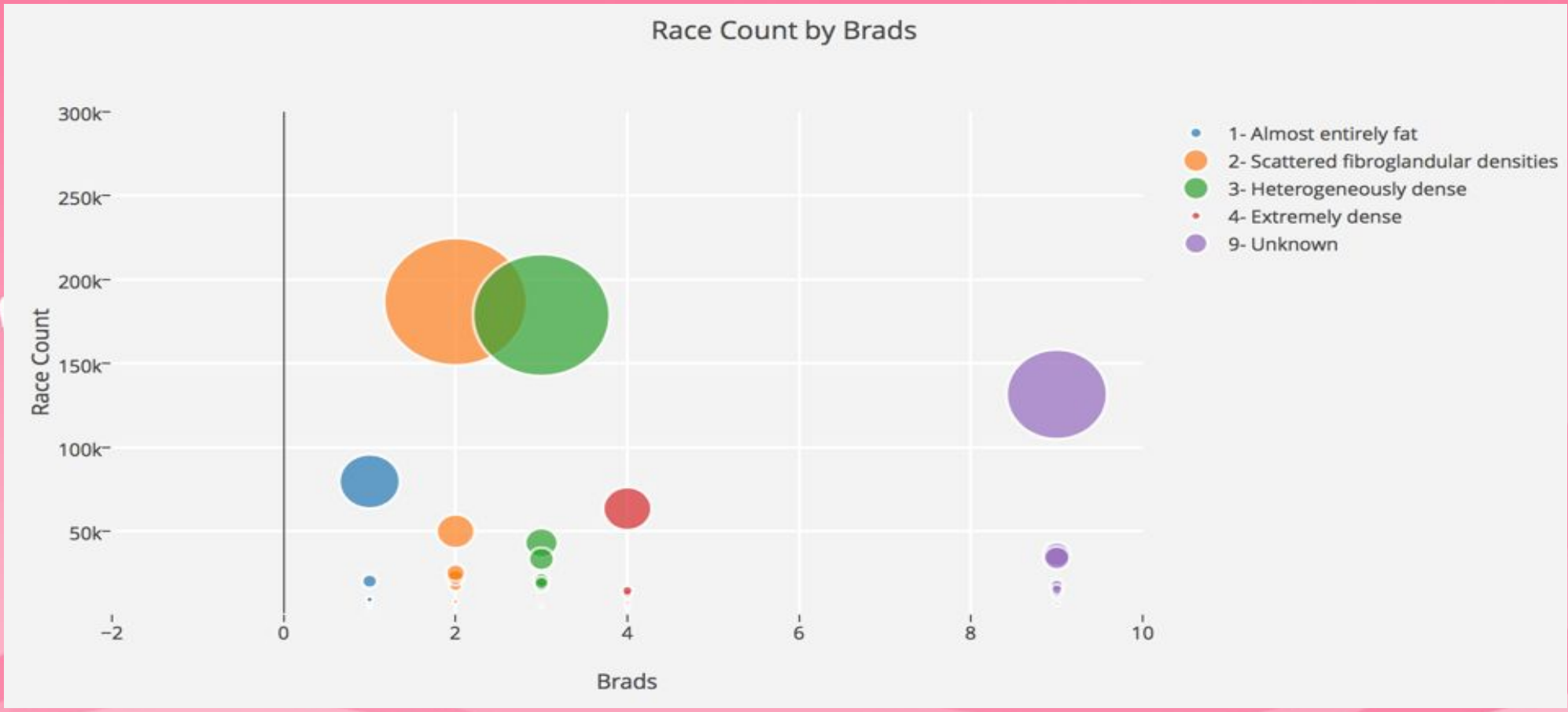

Breast Cancer Data Population by Race
1- Non-Hispanic white
55.92%
5- Hispanic
13.93%
3- Asian/Pacific Islander
10.10%
9- Unknown
6.60%
2- Non-Hispanic black
6.40%
6- Other/mixed
4.96%
4- Native American
2.10%
0
10
20
30
40
50
60

Age Factor


Breast Cancer Data Population by Age
Ordered by Percentage
Ordered by age group
6- Age 50-54
5- Age 45-49
7- Age 55-59
4- Age 40-44
8- Age 60-64
9- Age 65-69
10- Age 70-74
11- Age 75-79
12- Age 80-84
3- Age 35-39
13- Age >85
2- Age 30-34
1- Age 18-29
17.75%
14.19%
11.64%
10.27%
10.19%
8.74%
7.52%
6.38%
4.57%
3.98%
2.79%
1.46%
.52%
0
5
10
15
20
0
5
10
15
1- Age 18-29
2- Age 30-34
3- Age 35-39
4- Age 40-44
5- Age 45-49
6- Age 50-54
7- Age 55-59
8- Age 60-64
9- Age 65-69
10- Age 70-74
11- Age 75-79
12- Age 80-84
13- Age >85


Relationship between patient breast cancer and first-degree relative cancer history
600k
500k
400k
300k
200k
100k
0
count
yes_relative_cancer_history
patient_yes_cancer_history
no_relative_cancer_history
patient_no_cancer_history
1- Almost entirely fat
2- Scattered fibroglandular densities
3- Heterogeneously dense
4- Extremely dense
9- Unknown
birads_breast_density

Race within BRADS



Modeling - LOGISTIC REGRESSION

Features and target

Features:

- age_menarche - Age (years) at menarche
- age_group_5_years - Age (years) in 5 year groups
- race_eth - Race/ethnicity
- first_degree_hx - History of breast cancer in a first degree relative
- breast_cancer_history - Prior breast cancer diagnosis
- age_first_birth - Age (years) at first birth
- current_hrt - Use of hormone replacement therapy
- menopaus - Menopausal status
- bmi_group - Body mass index

Target:

Results based on BI-RADS breast density:

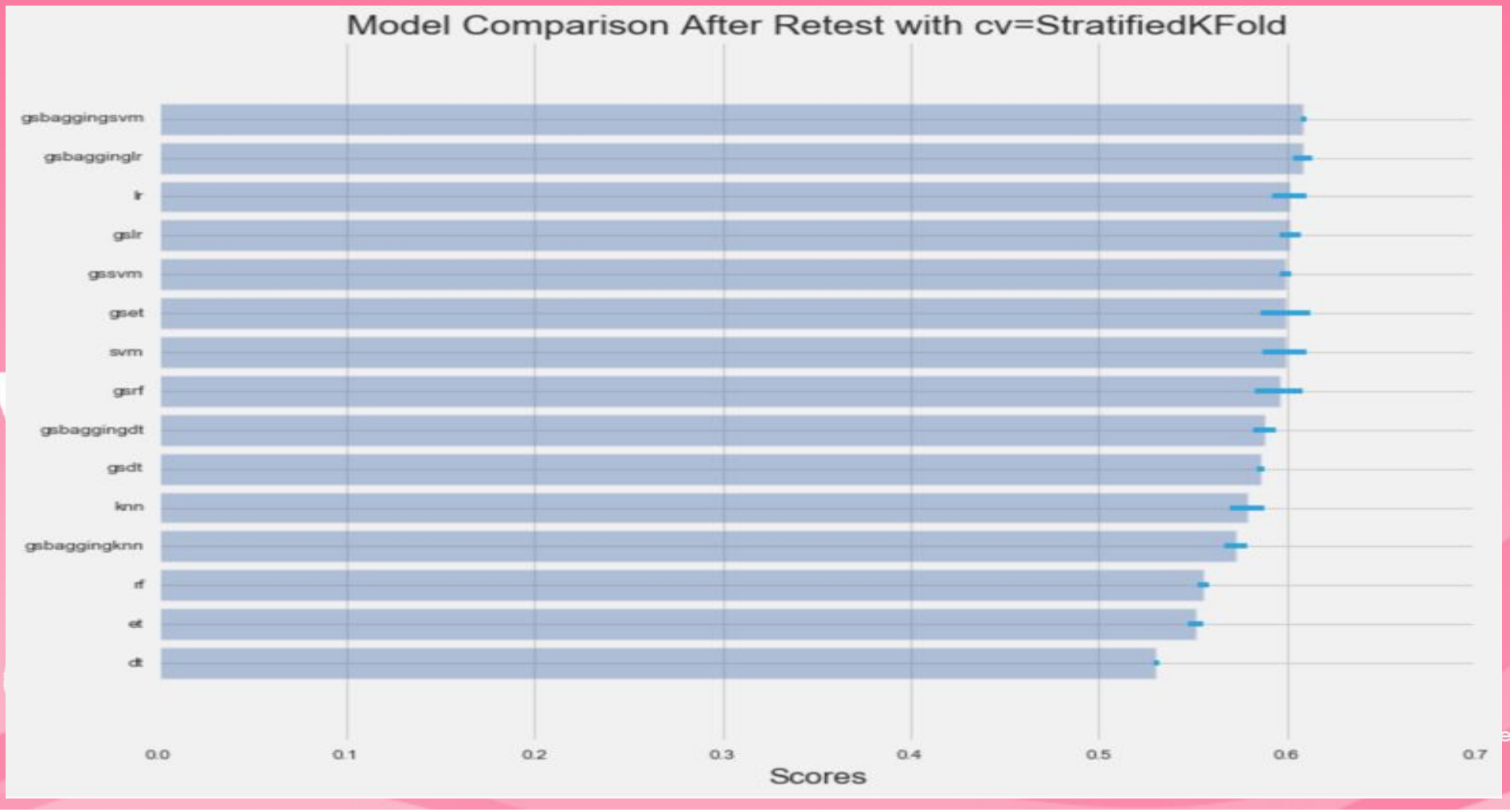
- 0 - Negative - Low risk
- 1 - Positive - High Risk

Modeling - LOGISTIC REGRESSION

I used a logistic regression model and compared it with the classifiers below:

	Model	Mean	STD
1	Decision Tree	0.544604	0.008135
2	Extra Trees	0.560403	0.005835
3	Random Forest	0.5616	0.010047
4	GridsearchCV bagging + knn	0.582396	0.008304
5	GridsearchCV dt	0.589392	0.015879
6	knn	0.592803	0.009506
7	svm - Support Vector Machine	0.597604	0.019189
8	GridsearchCV Bagging svm	0.604404	0.007468
9	GridsearchCV Bagging Decision Tree	0.610404	0.011112
10	GridsearchCV Random Forest	0.611399	0.004516
11	GridsearchCV svm	0.611803	0.004967
12	GridsearchCV Extra Trees	0.614007	0.013798
13	GridsearchCV Logistic Regression	0.614197	0.005397
14	Logistic Regression	0.616597	0.006232
15	GridsearchCV Bagging Logistic Regression	0.6232	0.0086

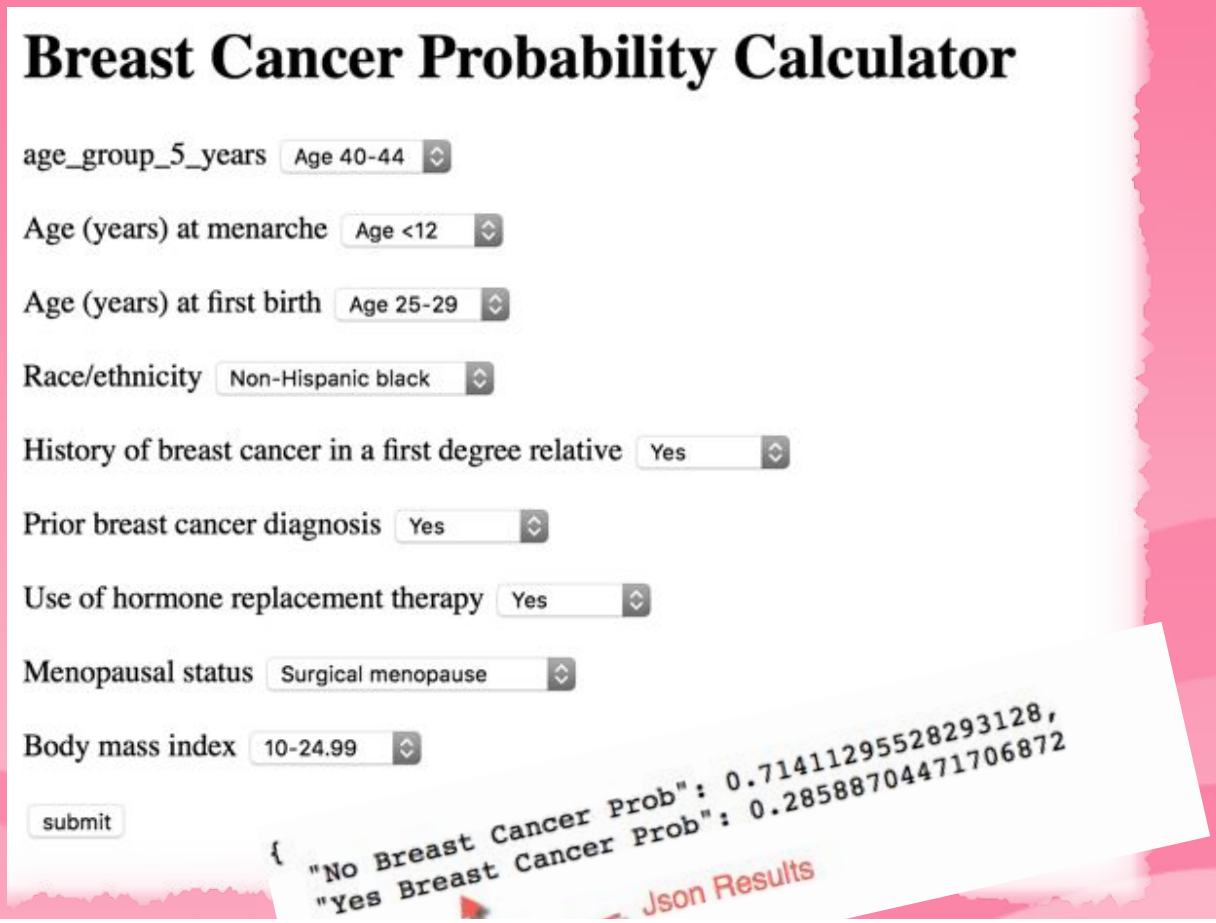
Model Comparison



Breast Cancer Probability Calculator

Logistic Regression Model

Probability Prediction



Risks & Assumptions

- The risk of obtaining accurate relative results from predicting based on factors, such as demographics, reproductive history, medications, genetic factors , and clinical and biologic markers (e.g., body mass index).

- Missing or biased data has to be taken into account when reviewing the results of the present study

Results and Conclusion

- Developed a prognostication model for early breast cancer.
- The performance and score have been compared among several classifiers.
- The best performer was GridsearchCV Bagging Logistic Regression.
- Picked the Logistic regression because of its simplicity.
- Being able to predict breast cancer outcomes more accurately would help physicians make informed decisions regarding the potential necessity of research and treatment in women patients.
- It will also contribute to raise awareness and funds to help women to reduce their risk of breast cancer

Questions???