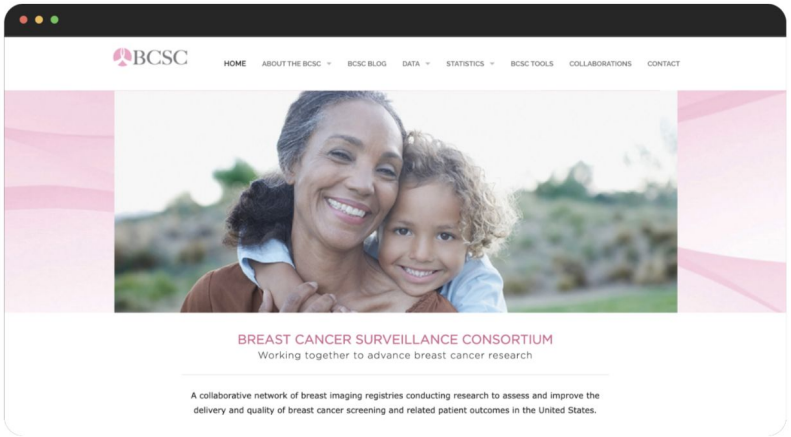
Finding the best model for Breast Cancer Classification

By: Aarav Desai, Sanjay Subramaniam, Tanmay Grandhisiri, Runze Shao, Shahab Khorasanizadeh, Justin Estes

Dataset and Background Information

Breast Cancer Surveillance Consortium



Features explored:

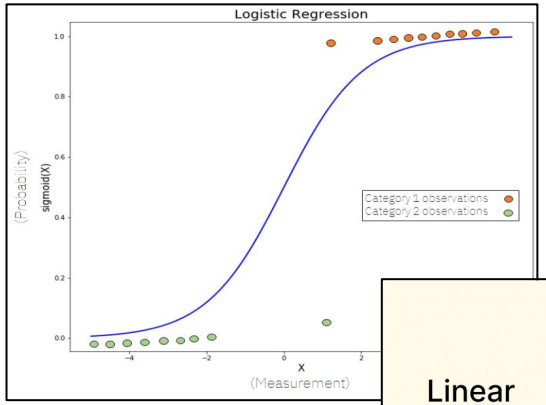
- Menopause
- Age Group
- Breast Density
- Race
- Hispanic
- BMI
- Diagnosis of invasive breast cancer
- No of relatives
- Previous breast procedure
- Result of last mammogram
- Surgical menopause
- Hormone therapy
- Age at first birth

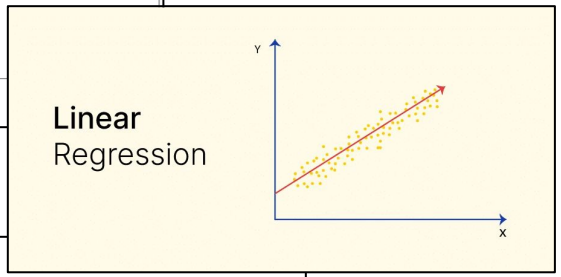
Questions we set out to answer?

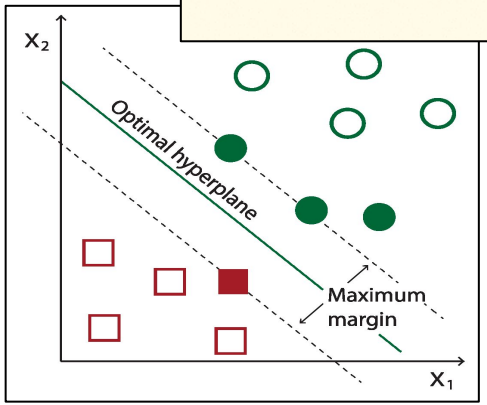
- We wanted to see if given a dataset of women patients, if we could find the best model in order to predict whether a person would have breast cancer depending on certain factors.
- With the results we get from our project we could then see which factors have the highest correlation with a positive breast cancer diagnosis.

Models

- Multiple Linear Regression
- Logistic Regression
- Support Vector Machines (SVM)



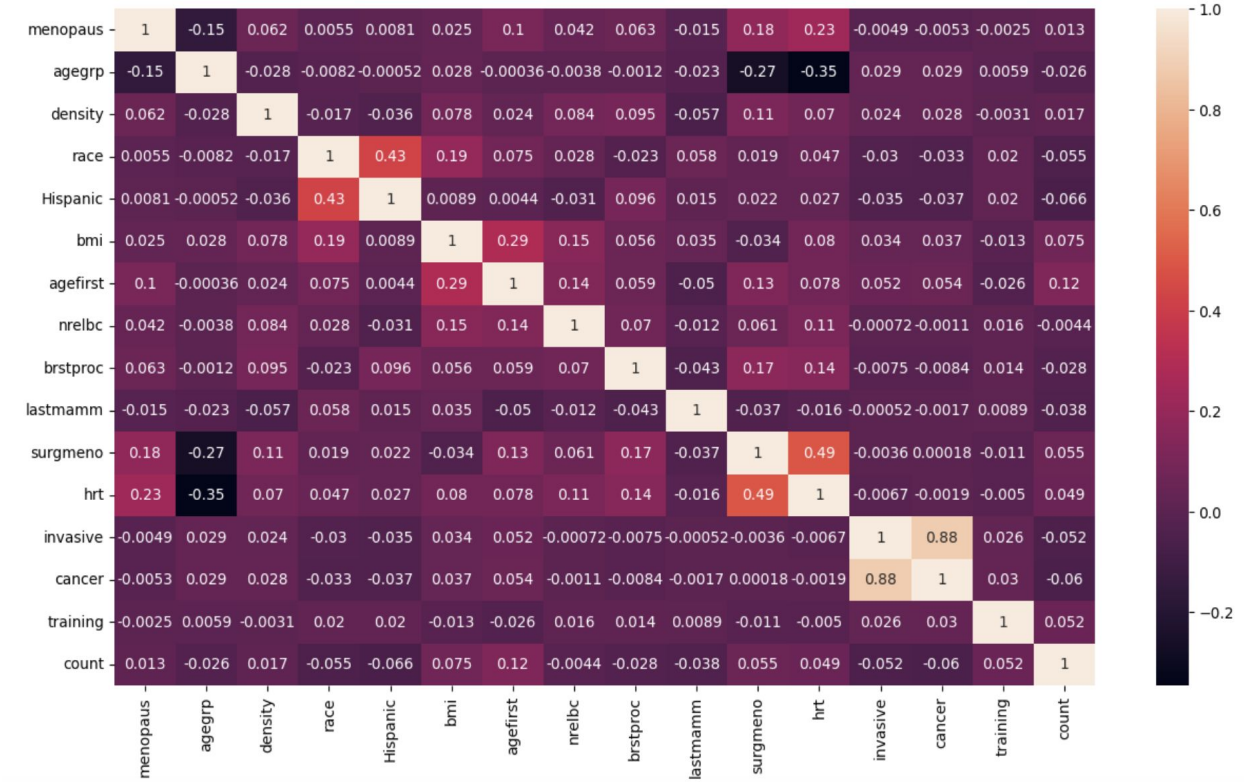




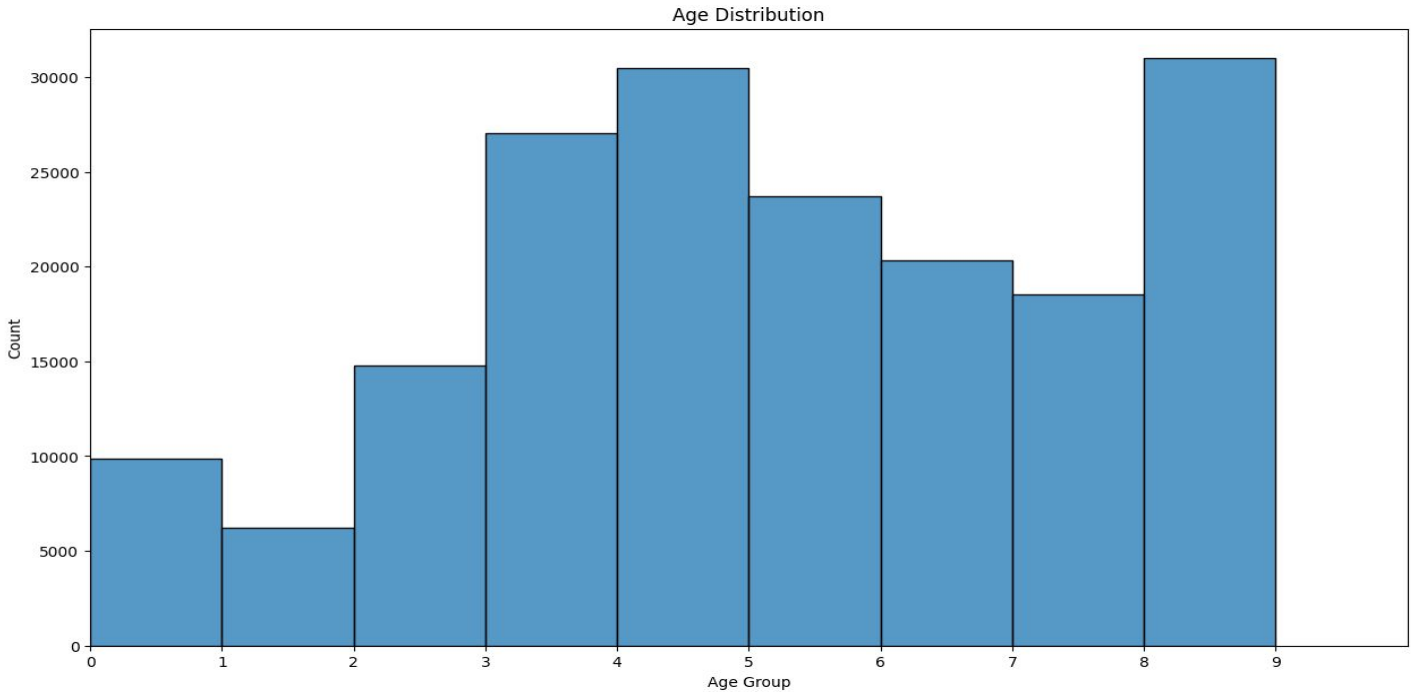
Computational Techniques

- Data Visualization: Matplotlib and Seaborn
- Data Analysis: Numpy and Pandas
- Machine Learning: Scikit-learn
- Regression Models: Statsmodel

Correlation Matrix (Heatmap)

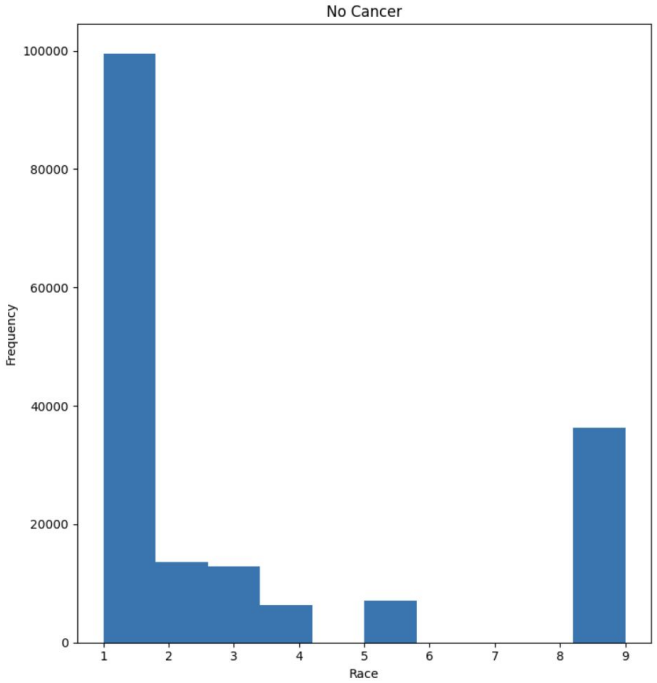


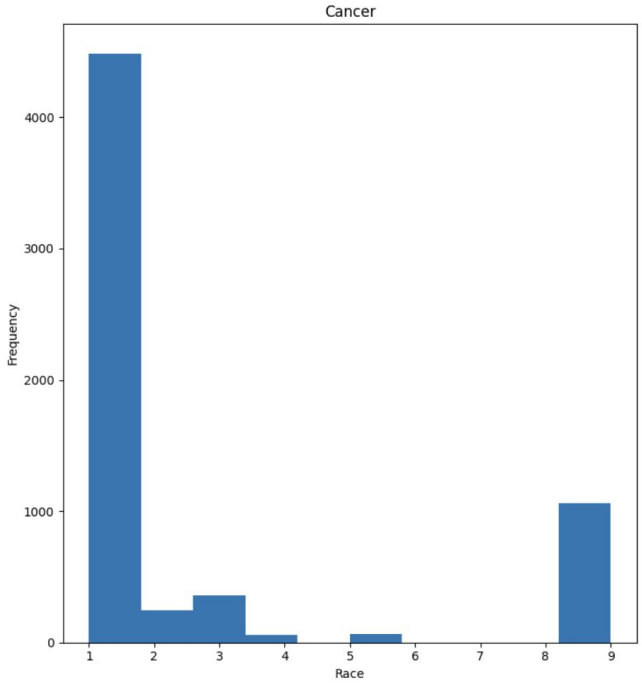
Histogram showing Age Distribution



Age group: 1 = 35-39; 2 = 40-44; 3 = 45-49; 4 = 50-54; 5 = 55-59; 6 = 60-64; 7 = 65-69; 8 = 70-74; 9 = 75-79; 10 = 80-84

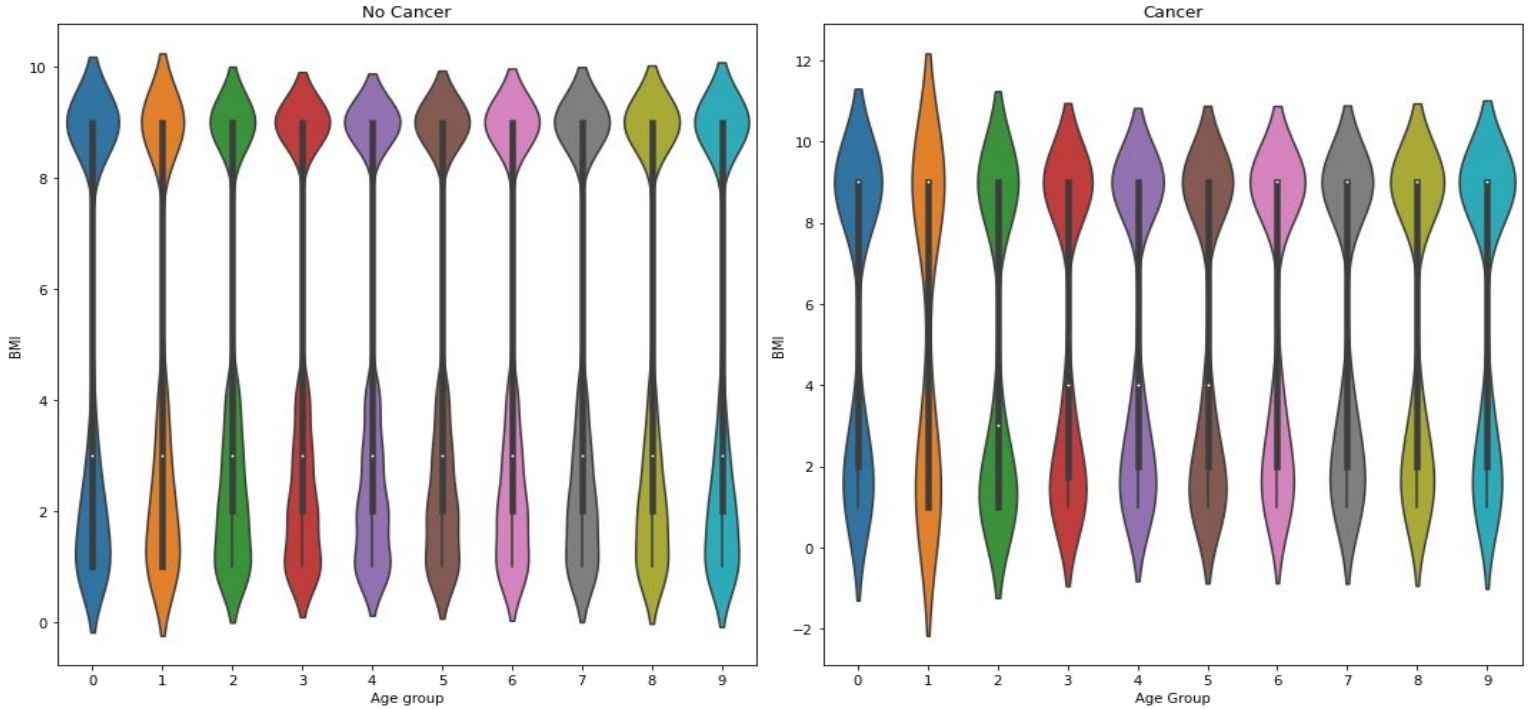
Frequency of Cancer and Non-Cancer patients for different races





1 = white; 2 = Asian/Pacific Islander; 3 = black; 4 = Native American; 5 = other/mixed; 9 = unknown

Violin Plot (BMI vs Age)



Body mass index: 1 = 10-24.99; 2 = 25-29.99; 3 = 30-34.99; 4 = 35 or more; 9 = unknown

Age group: 1 = 35-39; 2 = 40-44; 3 = 45-49; 4 = 50-54; 5 = 55-59; 6 = 60-64; 7 = 65-69; 8 = 70-74; 9 = 75-79; 10 = 80-84

Multiple Linear Regression

- R-squared: 0.785
- Adj R-squared: 0.785

OLS Regression Results

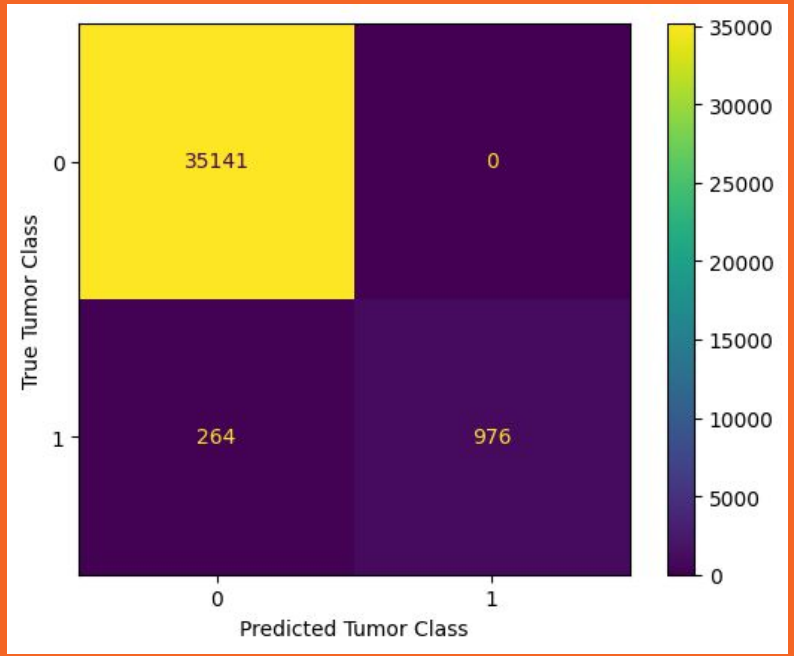
Dep. Variable:	cancer	R-squared (uncentered):	0.785
Model:	OLS	Adj. R-squared (uncentered):	0.785
Method:	Least Squares	F-statistic:	5.104e+04
Date:	Tue, 18 Apr 2023	Prob (F-statistic):	0.00
Time:	16:10:01	Log-Likelihood:	1.8787e+05
No. Observations:	181903	AIC:	-3.757e+05
Df Residuals:	181890	BIC:	-3.756e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
menopaus	-0.0002	0.000	-1.738	0.082	-0.000	2.27e-05
agegrp	0.0006	6.59e-05	9.524	0.000	0.000	0.001
density	0.0005	6.54e-05	7.362	0.000	0.000	0.001
race	-0.0004	7.25e-05	-5.435	0.000	-0.001	-0.000
Hispanic	-0.0001	5.6e-05	-2.097	0.036	-0.000	-7.67e-06
bmi	0.0004	6.11e-05	6.446	0.000	0.000	0.001
agefirst	0.0004	5.3e-05	6.863	0.000	0.000	0.000
nrelbc	-0.0002	6.84e-05	-2.842	0.004	-0.000	-6.04e-05
brstproc	-0.0002	7.29e-05	-3.411	0.001	-0.000	-0.000
lastmamm	3.813e-05	4.53e-05	0.842	0.400	-5.06e-05	0.000
surgmeno	0.0002	5.42e-05	2.820	0.005	4.66e-05	0.000

Multiple linear regression is a regression model that estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line.

Logistic Regression

- Accuracy - 99.25%
- Standard deviation - 0.07%

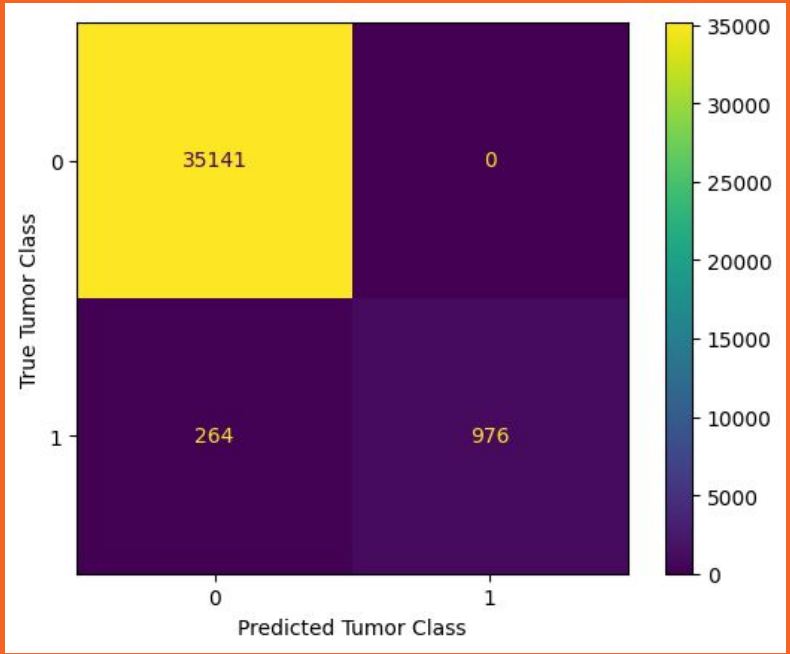


Support Vector Machine (SVM)

- Precision - 0.99
- Recall - 1.00
- F1 Score - 1.00
- Accuracy - 99.27%

Mean accuracy of 99.25% was obtained after K-fold cross validation showing that the SVM is not overfitting the data

Same confusion matrix results as logistic regression



Answers to our question

- Cancer feature (0-1) and cancer invasive (highest correlation from heatmap) are mathematically correlated, given R-Square = 0.777
- All the features with “cancer” are mathematically correlated, by applying multiple regression, given R-Square = 0.785
- Based on the logistic regression, prediction accuracy is 99.25%.
- Based on the logistic regression, the misclassification rate is 0.72% (Machine error can cause slight difference Accuracy -> **100%-0.72% ≠ 99.25%**)
- SVM is the best model with 99.27% accuracy

Difficulties or complications

- Long run time of the RBF kernel vs Linear kernel. Took too long to see output.
- Tried to use a multilayer perceptron, but there was too much data for the model to look at so we went with logistic regression instead.
- Linear model and regression gave the same results, so we removed the model
- In our data, we had a lot of values in important column that were not available (Ex: BMI = 9 -> Unknown)

References

"Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C). You can learn more about the BCSC at: http://www.bcsc-research.org/."

Bevans, Rebecca. “Multiple Linear Regression: A Quick Guide (Examples).” *Scribbr*, 15 Nov. 2022, https://www.scribbr.com/statistics/multiple-linear-regression/.

Pramoditha, Rukshan. “K-Fold Cross-Validation Explained in Plain English.” *Medium*, Towards Data Science, 20 Dec. 2020, https://towardsdatascience.com/k-fold-cross-validation-explained-in-plain-english-659e33c0bc0.