Machine Learning Applications to Triple-Negative Breast Cancer Prognosis and Prediction

Wei-Hua Hsu (Wafer) | Data Science, American University

INTRODUCTION

- Triple-Negative Breast Cancer (TNBC) is the worse cancer that tests negative for estrogen receptors (ER), progesterone receptors (PR), and excess HER2 protein.
- 42.7% of TNBC patients die in 2 years.[1]
- Age effect in breast cancer is significant while it is not clear in TNBC patients. Also, TNBC does not respond to hormonal therapy medicines.

CONTRIBUTION

- We firstly select the variables based on survival analysis results[1]. Besides, we use Feature Selection to obtain a subset model that is interpretable and predictable in enhancing the predicting accuracy.
- Supervised and unsupervised learning methods are applied, which intend to correctly predict the labels of two years survivability based on the characteristics of the patient.

DATA & FEATURES

The proportion of the 2 years survival (Y) is 43% (<=2 years) with 57% (>2 years).

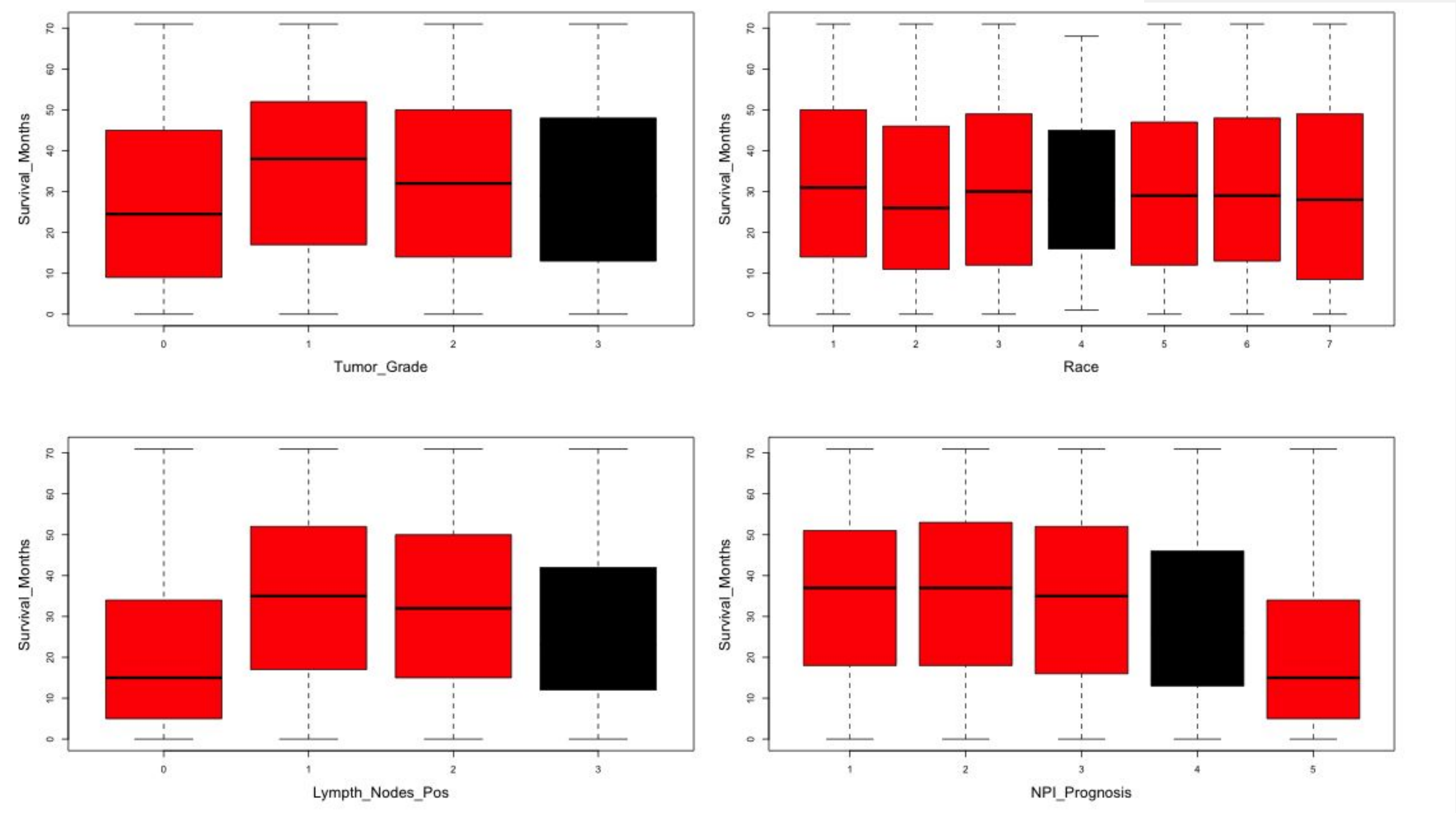
TNBC Dataset[1]

- Race/Ethnicity
- Age groups
- Age at diagnosis
- Survival months
- 2/5 years survival
- Tumor Grade
- Causes of death
- Vital status
- Lymph nodes +
- Tumor size (cm)
- NPI prognosis
- Marital status

CLUSTERING

K-Means clustering is easily to be affected by the initial partition. The confusion matrix and accuracy rate given that how predicted labels are changing. However, the model with accuracy = 0.528 returns the centroids which are consistent to the true labels. The centroids for label 2 (survive <= 2 Years) can be found from the black boxplots. Which are Grade = 3 (worse), Lymph Nodes = 3 (worse), NPI = 4 (Poor), and Race = 4 (HBlack).

```
    pred             pred            pred             pred
true    1      2 true    1    2 true    1    2 true    1    2
   1 13133   176    1 7175 6134    1 5775 7534    1 7166 6143
   2  9737   190    2 4831 5096    2 4869 5058    2 4828 5099
[1] 0.5733775   [1] 0.5281029 [1] 0.4662162 [1] 0.5278447
```

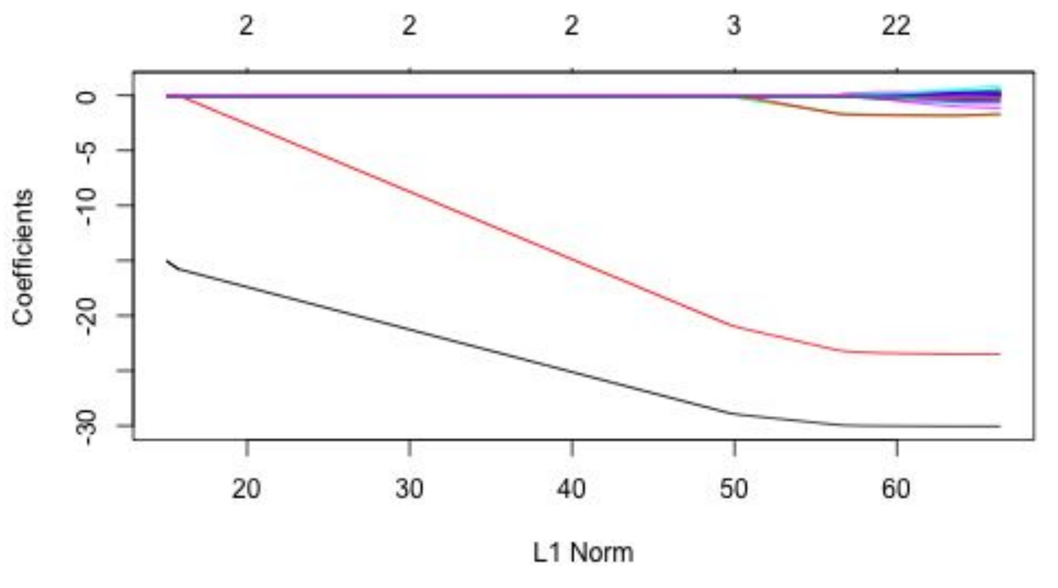


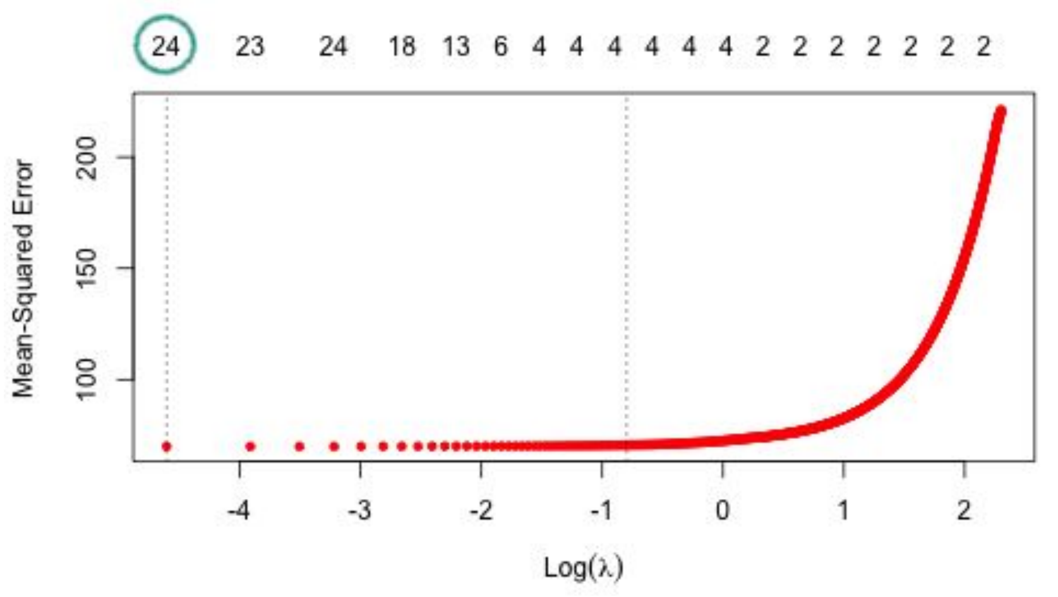
FEATURE SELECTION

Shrinkage method

The loss function in Lasso Regression is edited to minimize the model complexity by limiting the sum of the absolute values of the model coefficients (L1 Norm).

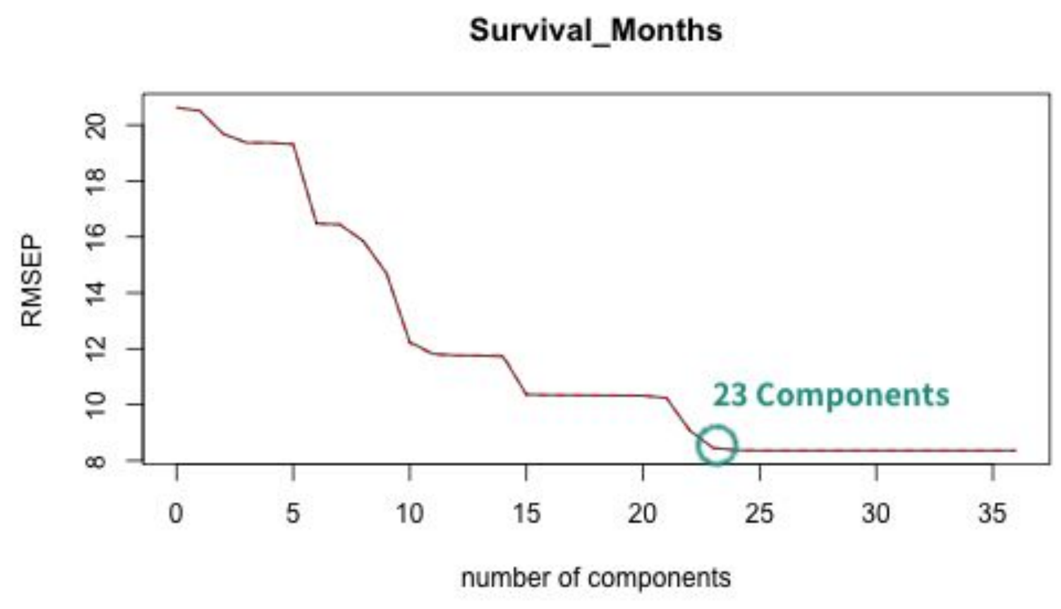
The second plot shows the best (minimal) CV λ = 0.01 that represents the least MSE. Which is the straight line across *log(λ) = -4.6* comes out to suggest 24 components.





Dimension Reduction

Obtain the first *P* principal components that account for 2 years survivability (Y). The validation plot below displays the *MSE = 8.46* in 23 components that explains 94.8% of the data.



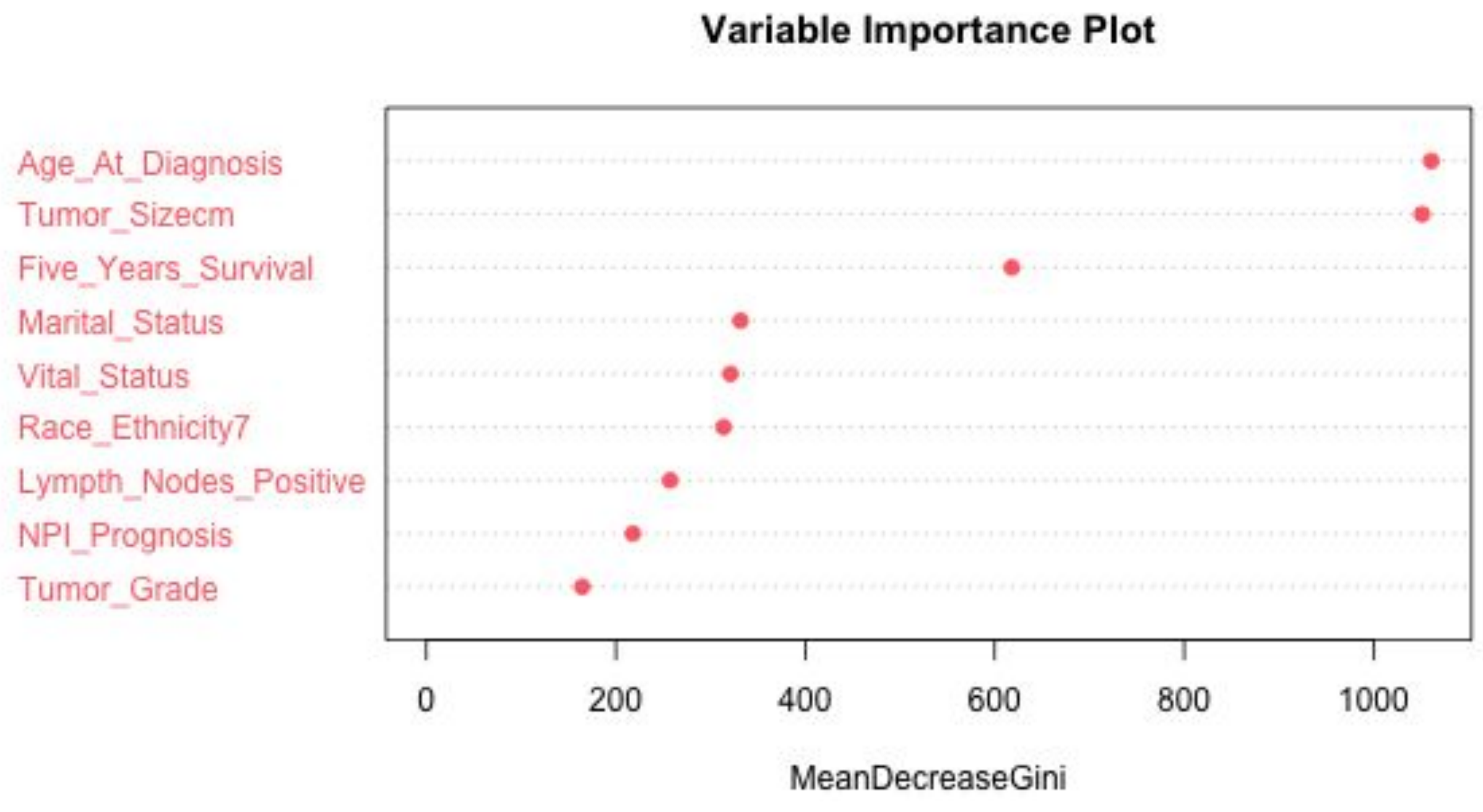
The adjusted R^2 is 83.23. The first *P* components obtained from the coefficient table included Race, Age, 2 years & 5 years survival, Vital status, Lymph nodes, Tumor grade & size, NPI prognosis, and Marital status.

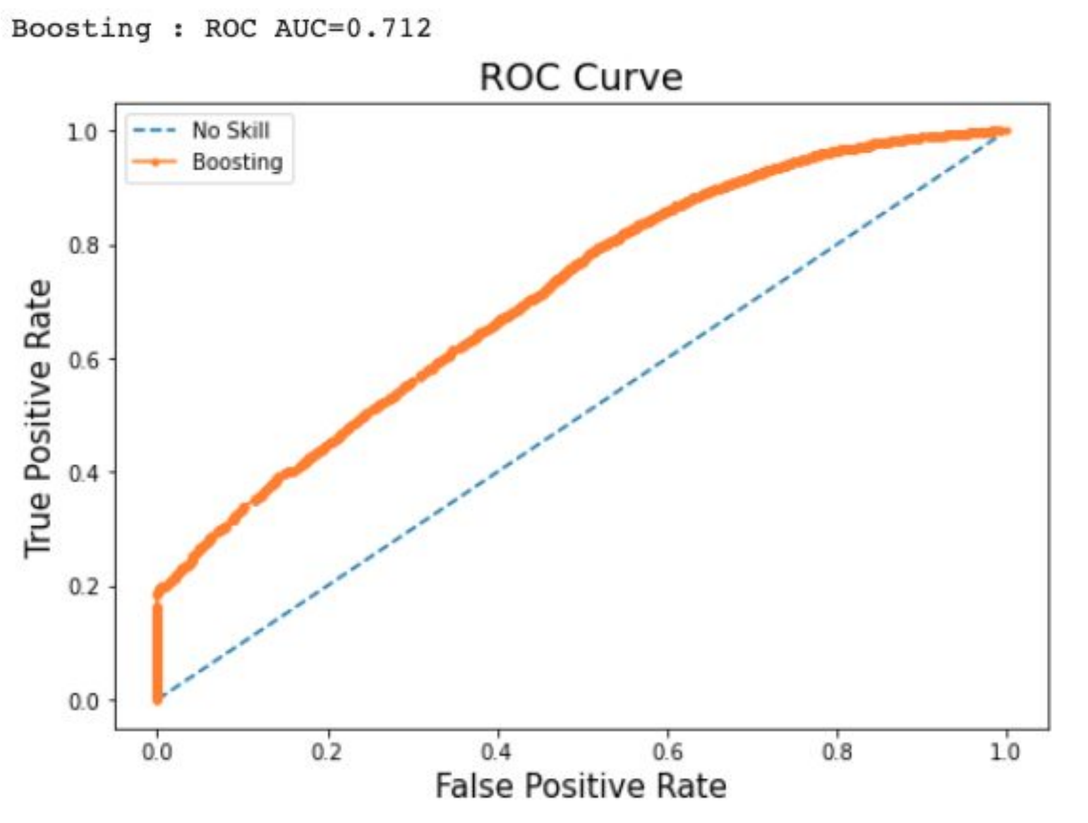
CLASSIFICATION

The table below illustrates the evaluations by eight classification methods. According to the table, boosting method achieves the highest accuracy and AUC score. Also, the ROC curve of Boosting is printed below.

The variable importance plot applies the random forest algorithm to illustrate how important the variable is in classifying the data. The chart displays that the `age at diagnosis` and `tumor size` are more important.

	AUC	F1-Scores	Accuracy
Naïve Bayes	0.644	0.7177	0.6448
Logistic Regression	0.652	0.7251	0.6478
SVM	0.500	*0.7283	0.5727
K-Nearest Neighbors	0.564	0.4901	0.5247
Boosting	*0.712	0.728	*0.657
Bagging	0.597	0.6319	0.5831
Random Forest	0.609	0.658	0.5928
Decision Trees	0.613	0.6733	0.6022
Mean	**0.6046**	**0.6688**	**0.6021**





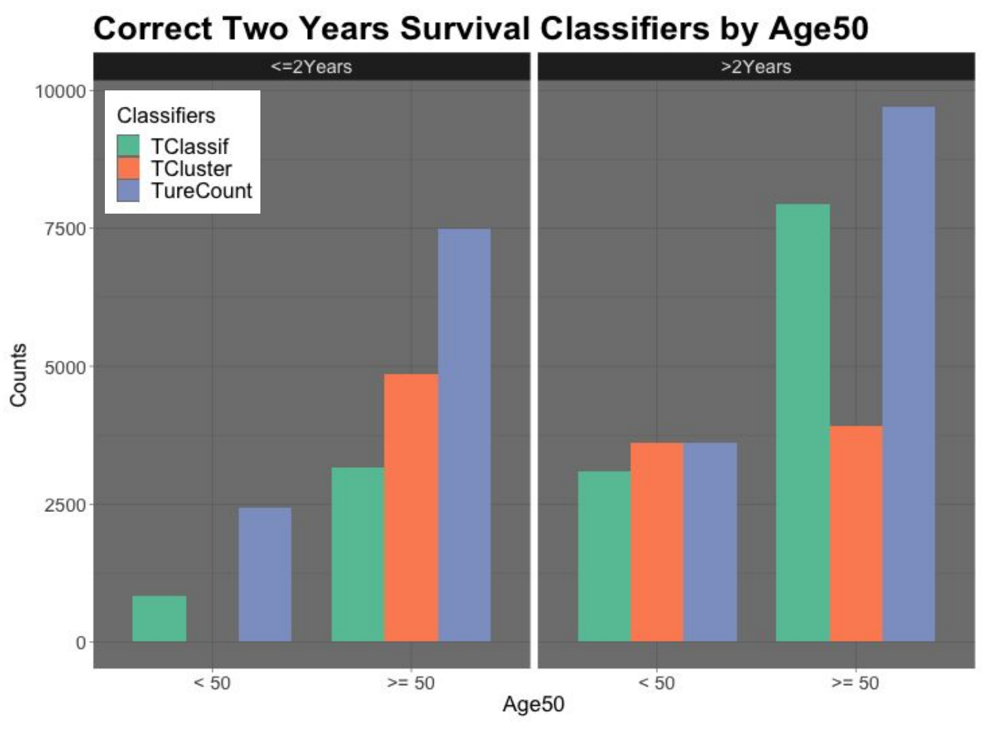
RESULTS & CONCLUSION

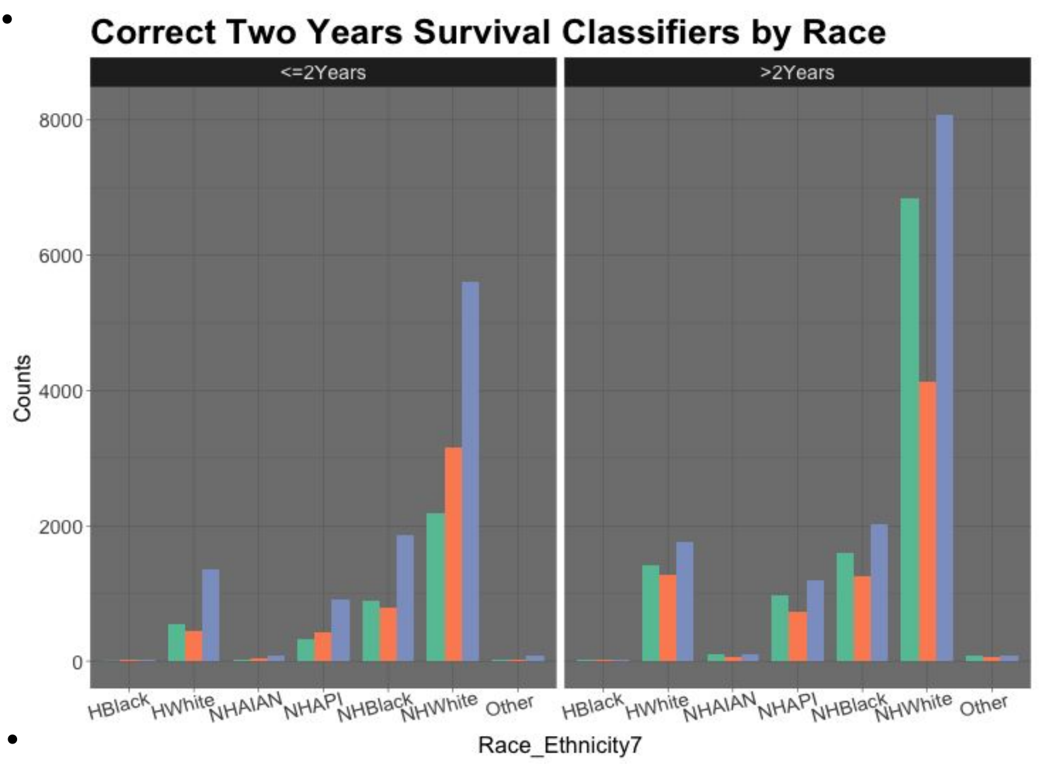
Results

- The bar charts compare the counts of correct predictions toward true labels (purple).
- Classification (green) does pretty well in predicting ≥ 2 years survival (right facet).
- Both methods can't predict < 2 years quite well, but clustering does better at more times.

Conclusion

- It is feasible to conduct TNBC prognosis by using machine learning techniques, but we need to do more experiments and capture more important features.
- Feature selection does help to omit less important variables.
- Classification does better in model prediction.





FUTURE WORK

TNBC Prognosis and Prediction

- Discover the characteristics and relationships of the patients in correct classification.

Increase Accuracy

- Hyper parameter tuning current model.
- Obtain more observations and variables. More important features are expected to improve the TNBC prognosis, such as obesity, cancer tissue, eating habits, or related disease/cancer history.

[1] Owrang, M., Kanaan, Y. M., & Dewitty JR., R. (2019). Ethnicity-related survival analysis of patients with triple-negative breast cancer. *EPiC Series in Computing, 58*, 236-246.

[2] Weigel, M. T., &amp; Dowsett, M. (2010). Current and emerging biomarkers in breast cancer: Prognosis and prediction. *Endocrine-Related Cancer, 17*(4). doi:https://doi.org/10.1677/ERC-10-0136

Machine Learning **Email:** wh0225a@american.edu **Website:** https://github.com/wafer110