

Bellevue University

DSC 680

Winter 2021

**Project Three:**

**Breast Cancer Prediction Based on Tumor Malignancy**

February 27, 2021

Jennifer Ruiz

**Abstract**

Breast cancer detection and early treatment are vital for positive patient outcomes. When detected early, tumor identification and removal can lead to decreased risk of regrowth and migration to other parts of the body. This hinges upon tumor malignancy results. By identifying key tumor characteristics that predict malignancy, that body of research grows, and patient outcomes improve.

In this project, the author chose to focus on tumor characteristics as predictors of malignancy in breast cancer. Using machine learning methods, the author seeks to determine characteristics which influence malignant growth. Visualization methods will be used to explore the relationships between factors and identify potential areas for future research.

*Keywords:* Breast Cancer, Decision Tree, Random Forest, Python

**Introduction**

Breast cancer is the second leading disease-related cause of death among women across the United States and the leading disease-related cause of death for Hispanic and African American women. Early detection and treatment are vital for improved survival rates and decreased risk of metastatic growth. The American Cancer Society (2021) estimates that 281,000 new cases of invasive breast cancer will be diagnosed yearly. Tumor identification via characteristics and cell composition are necessary for proper treatment planning. According to the National Breast Cancer Foundation (2021), there are eleven types of breast cancer, each with its own distinct tumor characteristics. These factors make study of tumor masses crucial to the overall strategy for treatment and prevention of breast cancer.

The Wisconsin Breast Cancer dataset is published to further the study of tumor masses. The data is collected from imaging results of tumor masses suspected of malignancy. Additional data supplied details whether malignancy was determined. Using the breast cancer data, the author seeks to focus on understanding tumor characteristics and identifying features that lead to higher malignancy rates.

**Research Question**

This project endeavors to analyze the impact of multiple tumor characteristics and to predict increased risk of malignancy. To achieve this end, a comprehensive statistical and visualization analysis will be conducted.

Key Research Questions include:

1. What impact do size, texture, and compactness have in predicting breast cancer?
2. What factors lead to highest rates of malignancy?
3. What factors most contribute to a benign diagnosis?
4. What combination of factors are best for predicting malignancy?

**Background Information**

**Methods**

***Data Understanding***

When approaching this project, the author began by examining the individual attributes contained in the dataset. All data was coded numerically based on patient identification number. Attributes were visualized during Exploratory Analysis using Python (see appendix one) to gain insight into relationships within the data.

***Data Preparation***

Minimal data preparation was needed to prepare for exploratory analysis and modeling. One attribute, *Unnamed*, was removed from the dataset as it contained only null values. Additionally, the target variable was transformed from categorical to numeric to accommodate modeling.

***Modeling***

Data preparation allowed for multiple methods to be performed on the data. For all analyses, a test train split was employed on the dataset. Analysis was conducted using the sklearn libraries for Random Forest and Gradient Boost.

The author chose to a utilize random forest model to fully explore the story within the data. The model utilized all features within the dataset with the target variable being tumor malignancy. Then Principal Component Analysis was applied to the model and ten key features were identified. The random forest model was run again using the factors identified through PCA. Finally, a gradient boosting machine model was run to compare accuracy and recall between models.

**Results**

Exploratory analysis and the models created produced significant insights into tumor characteristics. Exploratory analysis revealed that benign tumors tended to have larger radii, a larger overall perimeter, and a larger overall area. It also revealed that tumor texture was similar in both benign and malignant tumors. The data also revealed four features that showed higher values in malignant tumors. These were the radius, texture, perimeter, and area mean. A final area of interest was that thirty-seven percent of observations within the dataset were diagnosed with cancer.

The first random forest model produced a model accuracy of 97.56 percent and a recall accuracy of 93.02 percent. After applying PCA to the model, the random forest model was run again. The addition of PCA increased model accuracy to 97.62 percent and recall accuracy to 95.35 percent. Finally, a gradient boosting machine model was run and produced a model accuracy of 97.37 percent after model optimization.

**Discussion**

The analysis produced some interesting insights regarding tumor characteristics of both benign and malignant tumors. The performance of the PCA with random forest model was similar to that of the gradient boosting machine. This indicates better prediction as the model learns the data. Future analysis will focus on applying these findings to new tumor images. Some surprising insights gained during analysis were that only ten of the thirty features within the dataset were ideal predictors of tumor malignancy.

**Conclusion**

As breast cancer continues to be a leading cause of death for women in the United States and around the world, it is important to thoroughly understand all factors that contribute to the presence of the disease. This analysis has shown several key features that can be applied to better predict malignancy from imaging. Applying this insight to diagnostics to improve patient outcomes.

**References**

American Cancer Society. (2021). How Common is Breast Cancer? Retrieved on February 27, 2021 from <https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html>

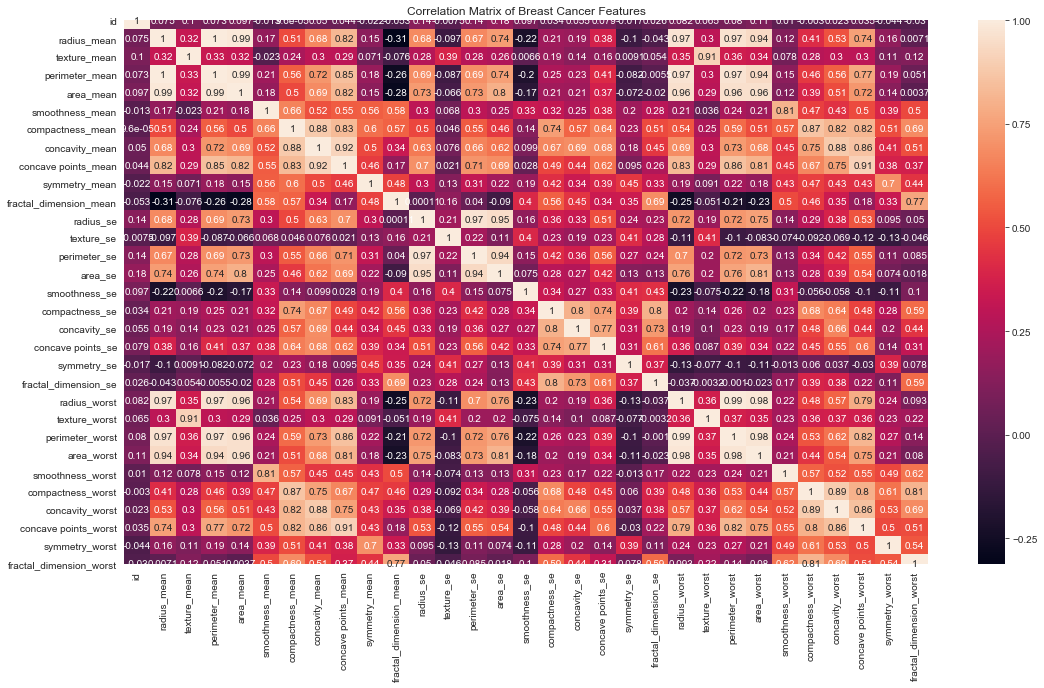
Kaggle. (n.d.). Breast Cancer Wisconsin Dataset. Retrieved on February 8, 2021 from <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>

National Breast Cancer Foundation. (2021). Types. Retrieved on February 27, 2021 from <https://www.nationalbreastcancer.org/types-of-breast-cancer/>

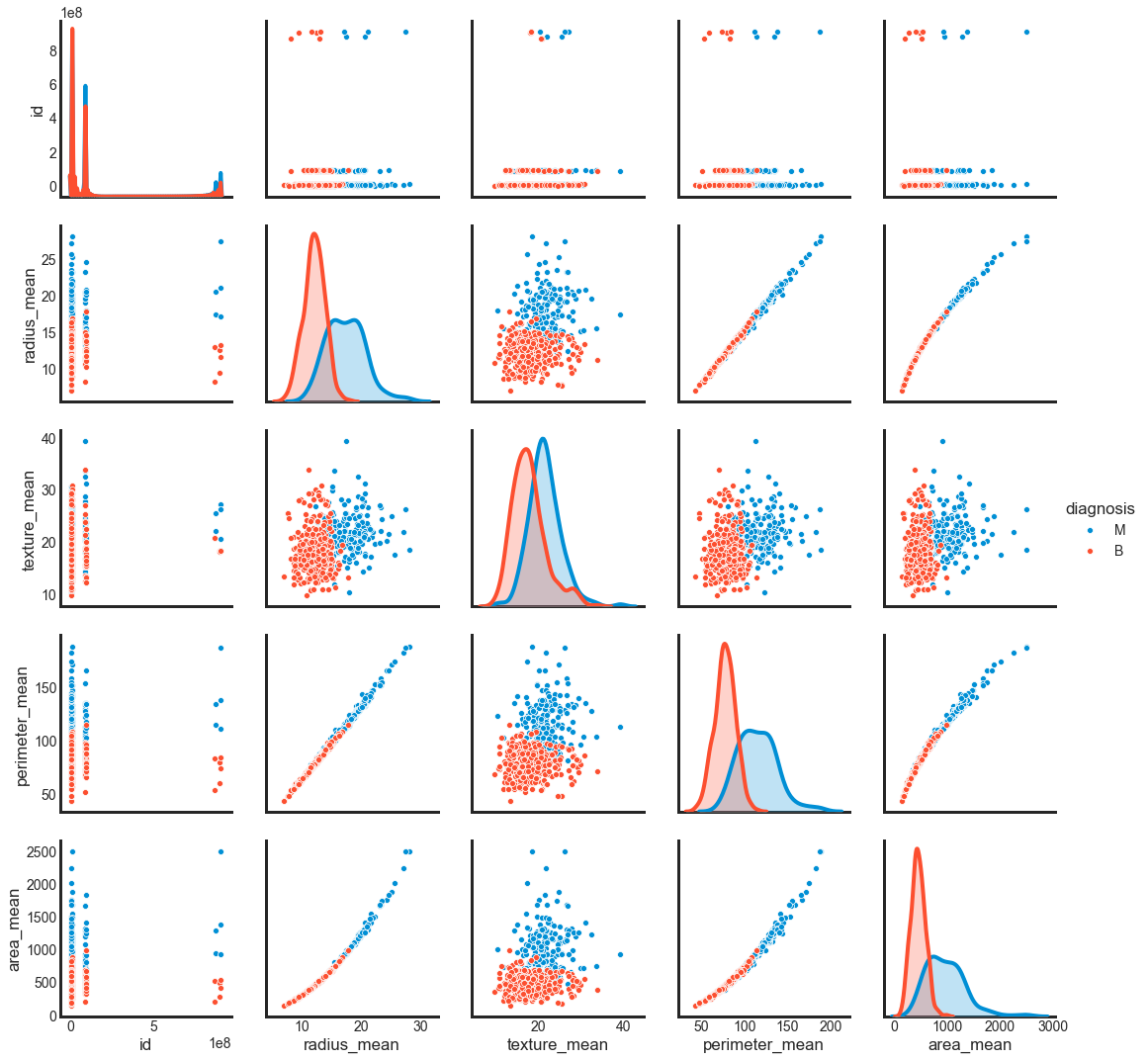
**Appendix A**

**Visuals created during exploratory analysis and modeling.**

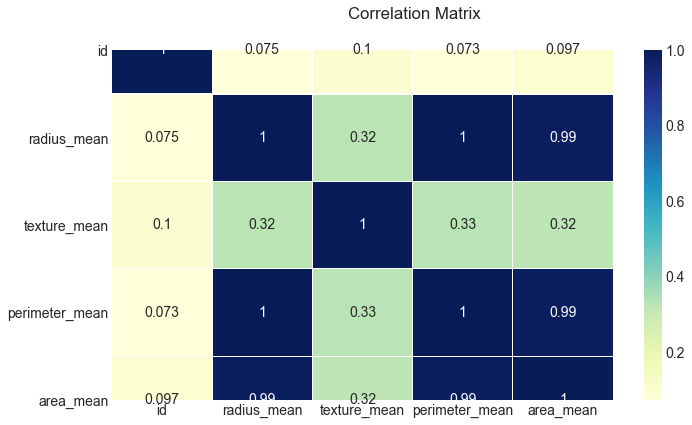
***Table One***

******

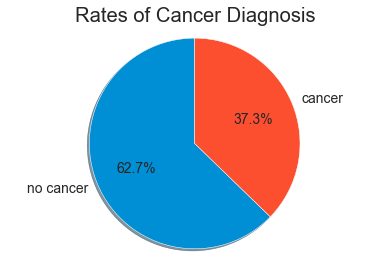
***Table Two***

****

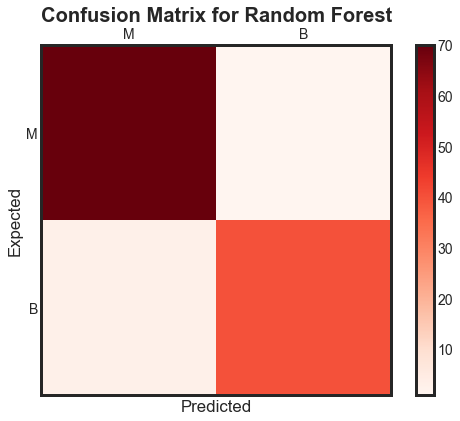
***Table Three***

******

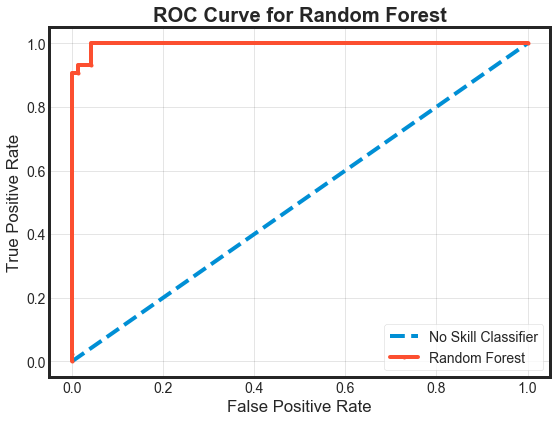
***Table Four***

****

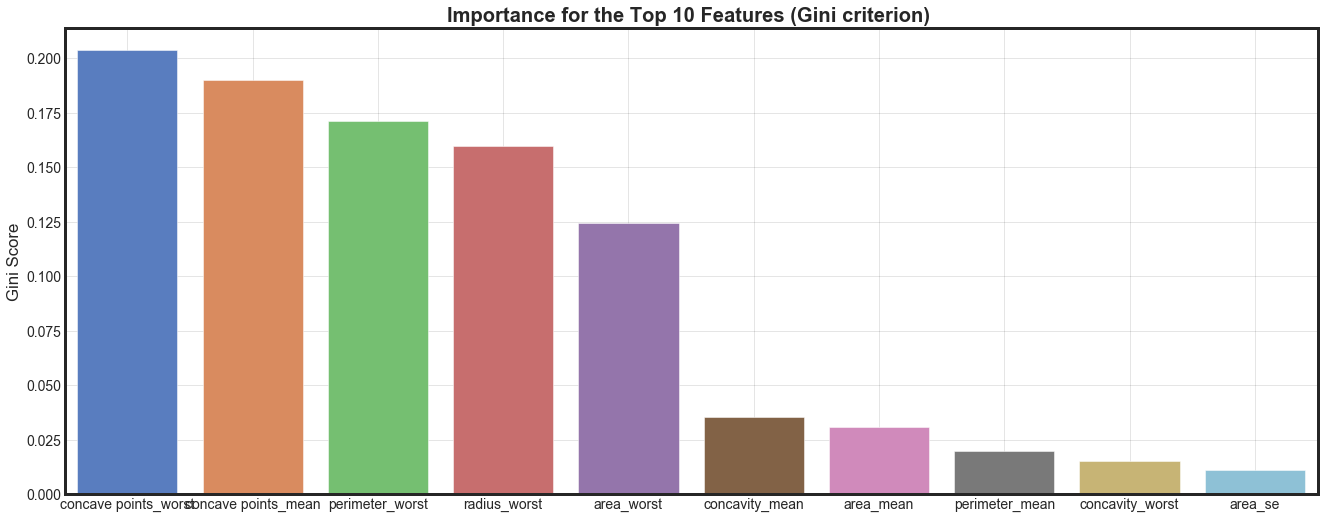
***Table Five***

****

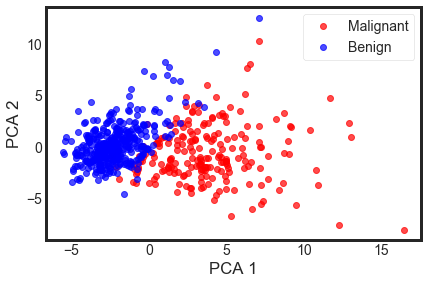
***Table Six***

******

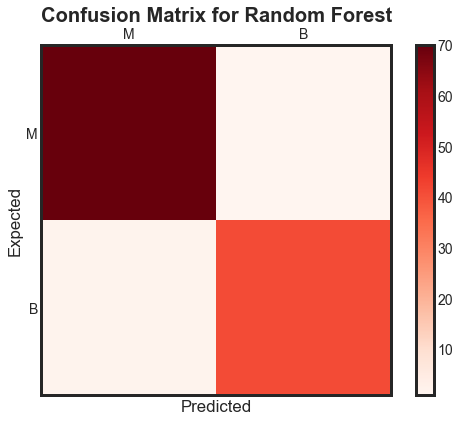
***Table Seven***

****

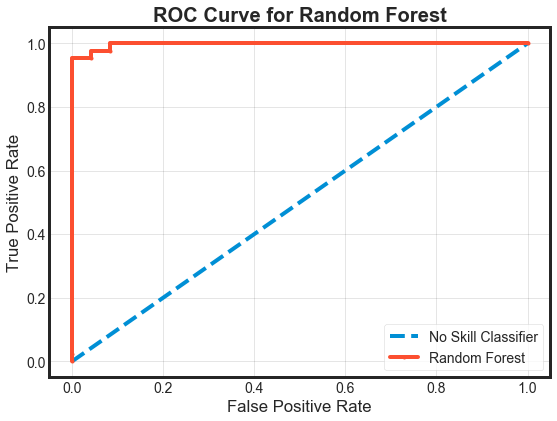
***Table Eight***

******

***Table Nine***

******

***Table Ten***

******