Pink Hope: The Fight Against Breast Cancer

Project Team 2:

Jeremiah Krick, Paul Thai, and Matthew Sexton

Shiley-Marcos School of Engineering, University of San Diego

AAI-500 - Probability and Statistics for Artificial Intelligence

Professor Dallin Munger

February 28, 2023

Abstract

The Breast Cancer Wisconsin dataset describes the characteristics of breast cancer cell nuclei. We analyzed the data of 569 cases, 30 features, and a diagnosis of malignant or benign. After exploratory data analysis, we selected the models random forest, XGBoost, and gradient boost to classify our data. Our goal was to determine which features have the largest impact in determining malignant and benign diagnosis.

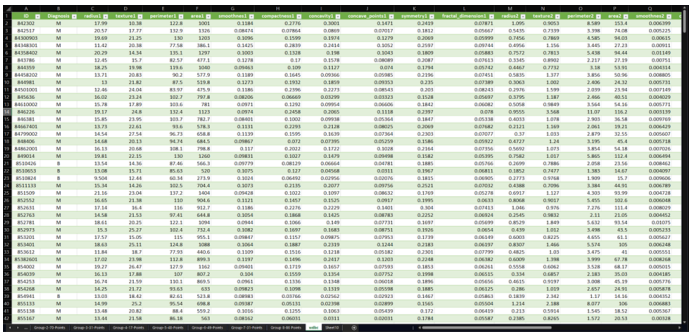
Keywords: breast cancer Wisconsin, exploratory data analysis (EDA), random forest, XGBoost, gradient boost.

Introduction

Accurately diagnosing breast cancer can be achieved by training a machine learning model to properly classify malignant and benign cells. By extracting and analyzing the breast cancer Wisconsin dataset we selected the following models Random Forest, XGBoost, and Gradient Boost to perform classification.

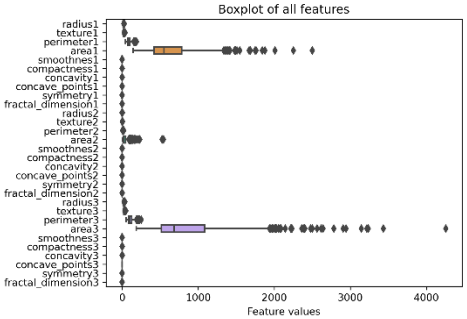
Data Cleaning/Preparation

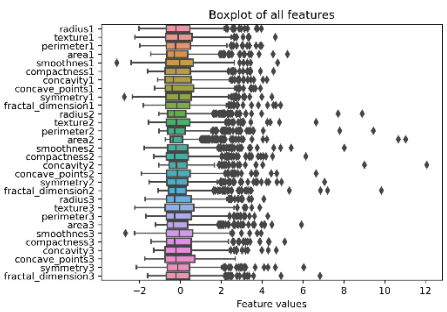
We prepared our data by loading it to an excel document for easy viewing. However, the wdbc.data file of 569 cases was ready to be used in the exploratory data analysis.



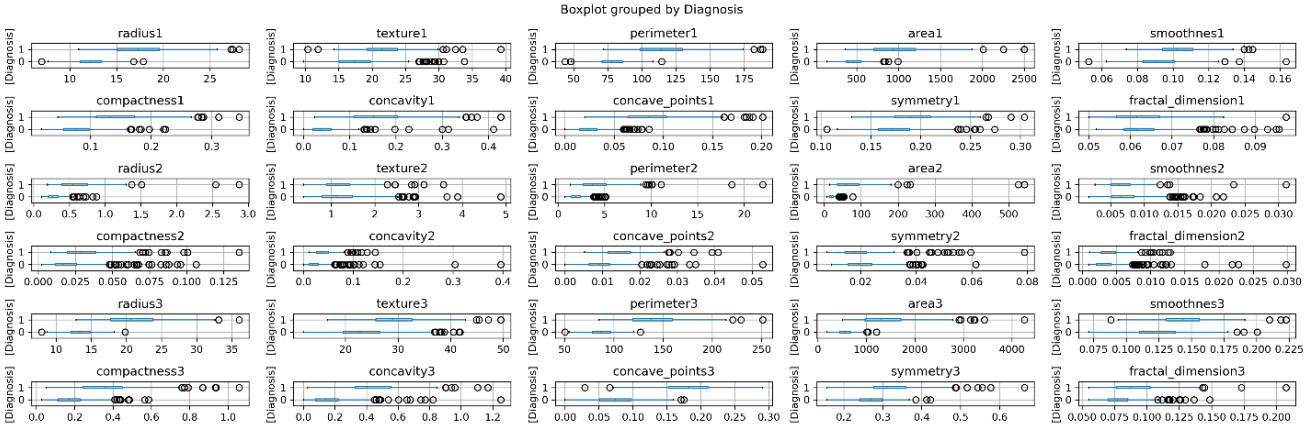
Exploratory Data Analysis

The initial step in the data analysis is to view all of the features to identify any abnormalities. The two obvious outliers in the box plot are area1 and area3, which annotate area mean and area worst, respectively. After performing a Z-score normalization on the 30 features, a new box plot displays the results.

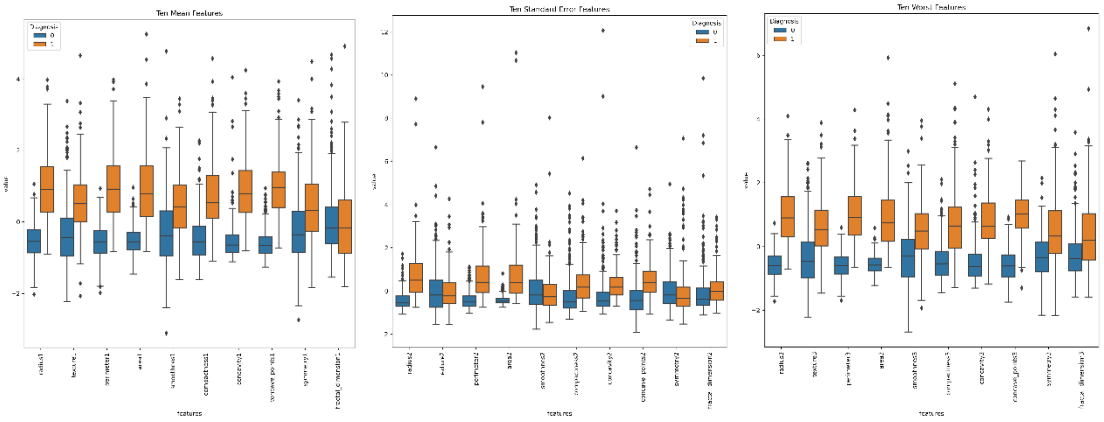




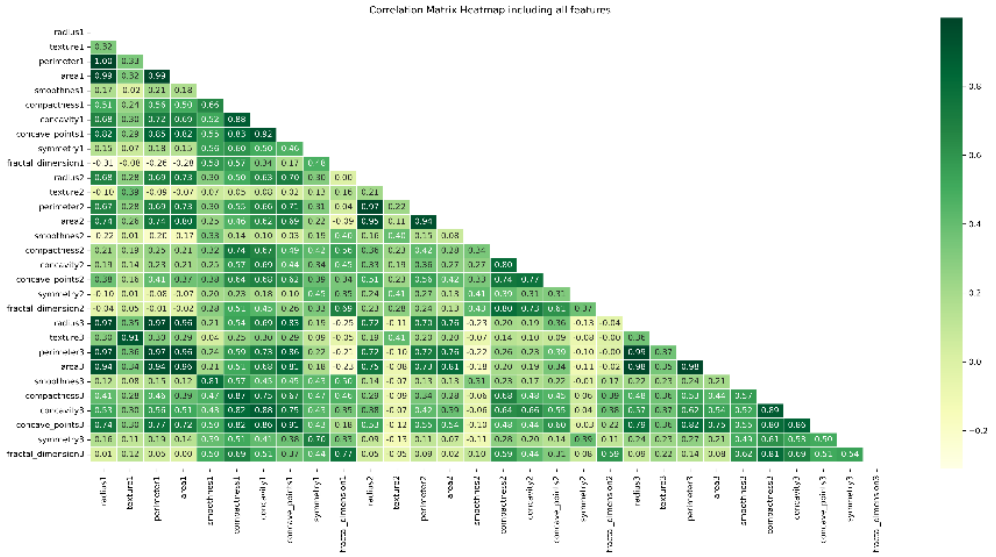
The next step was to view all 30 features split between malignant and benign classification. Below is an unnormalized representation using box plots.



The following box plots represent the ten z-score normalized features of mean, standard error, and worst values.



The correlation matrix heatmap easily displayed the positively correlated values. This information helps to better understand our data and shows that perimeter, area, and radius were correlated. This makes sense since these values are closely related when measuring nuclei.

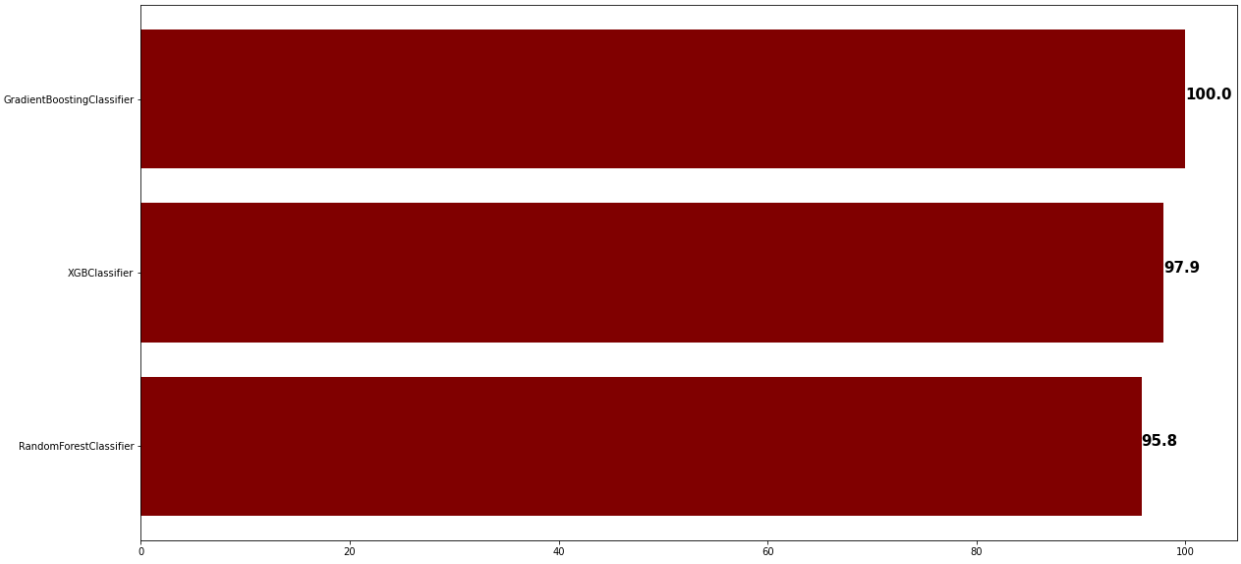


Model Selection

We opted to use the Random Forest Classification because it gives the highest average percentage in terms of accuracy and precision according to University of California Irvine Machine Learning Repository. We compare the Random Forest model against two other classification models, XGBoost and Gradient Boost to determine which features have the largest impact in diagnosing cancer cells.

Model Analysis

The Random Forest model achieved a 95.8% accuracy in predictions. Additionally, this model displayed the relative importance of model features. Compactness3 (worst), radius3 (worst), symmetry2 (SE), and texture3 (worst). The XGBoost model produced a 97.9% accuracy by itself, but when combined with the Gradient Boost model produced 100% accuracy. The top four important features of XGBoost were compactness3, compactness1, radius3, and texture1. Gradient Boost's top important features were symmetry2, radius3, compactness1, and compactness3.



Conclusion and Recommendations

Analyzing the three different results we concluded that compactness, radius, and symmetry all achieve the highest importance in each unique study. Random Forest classification had an overall accuracy of 95.8% which is on the lower end of the baseline model performance as shown in the UCI Breast Cancer Wisconsin Diagnostic dataset. XG boost, despite having a lower percentage accuracy on the UCI dataset, had a 97.9% accuracy on our model which closely resembles the same percentage of the dataset. Finally, gradient boosting at 100% accuracy. Having more models to compare would give us the ability to see whether different models would give us the same results as the previous. To our surprise, not only did this model give us 100% accuracy when we performed the calculations and graphed it, but it also validated that radius, symmetry, and compactness as the features with the highest importance. We suggest that using as many models as possible to study data would be the best way to achieve accuracy based on its versatile ability to use more collected data to compare and contrast results. We recommend future groups to also research and find other classification techniques to achieve better results.

Appendix

https://github.com/JeremyKrick/Team-Project-Statistics-for-AI

References

Learner CARES. (2020, July 21). *Building ML model to predict whether the cancer is benign or malignant on breast cancer wisconsin data set part 1*. Medium. https://medium.com/analytics-vidhya/building-ml-model-to-predict-whether-the-cancer-is-benign-or-malignant-on-breast-cancer-wisconsin-a09b6c32e7b8

R, A. (2018, July 31). *Applying random forest (classification) — Machine learning algorithm from scratch with real datasets*. Medium. https://medium.com/@ar.ingenious/applying-random-forest-classification-machine-learning-algorithm-from-scratch-with-real-24ff198a1c57

Sumbria, S. (2022, January 31). *Breast cancer wisconsin [diag] - Eda*. Kaggle. https://www.kaggle.com/code/shubamsumbria/breast-cancer-wisconsin-diag-eda

Wolberg, W., Street, W., & Mangasarian, O. (1995, October 31). *UC Irvine Machine Learning Repository*. https://archive-beta.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic