# Springboard Capstone 2 – Wisconsin Breast Cancer Prediction

Tzu-Ying Chen

# Problem Statement

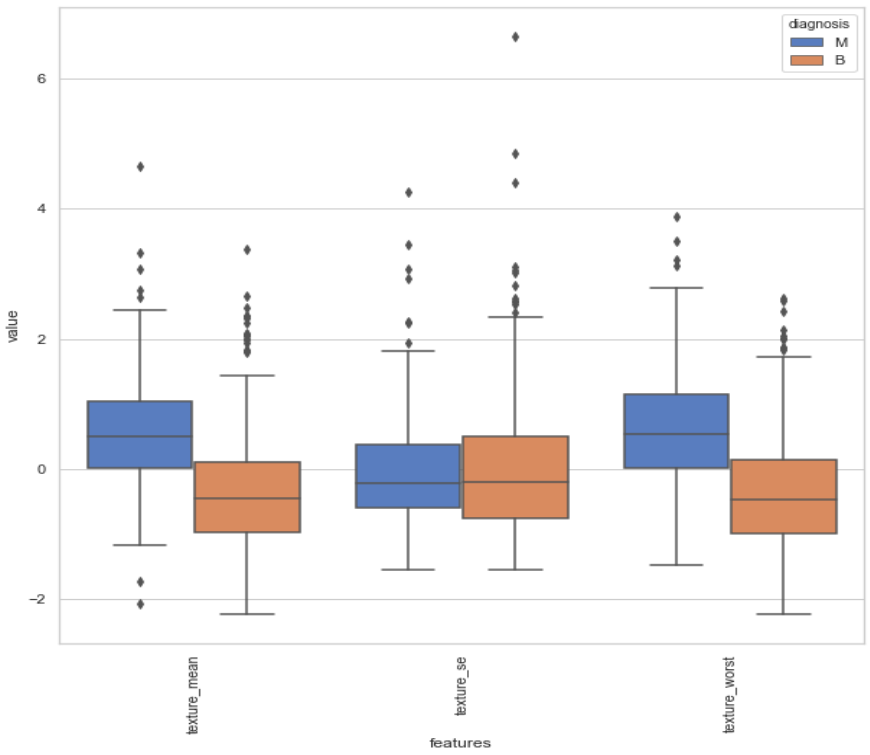
- There are numeric different characteristics of tumors that the breast cancer patients would have. The accuracy and efficiency of the diagnosis could significantly impact the patients' recovery and potentially decrease the complexity of the treatment. Hence, it is important to help doctors to identify the characteristics of the tumors that patients have in a timely manner in order to provide necessary healthcare.

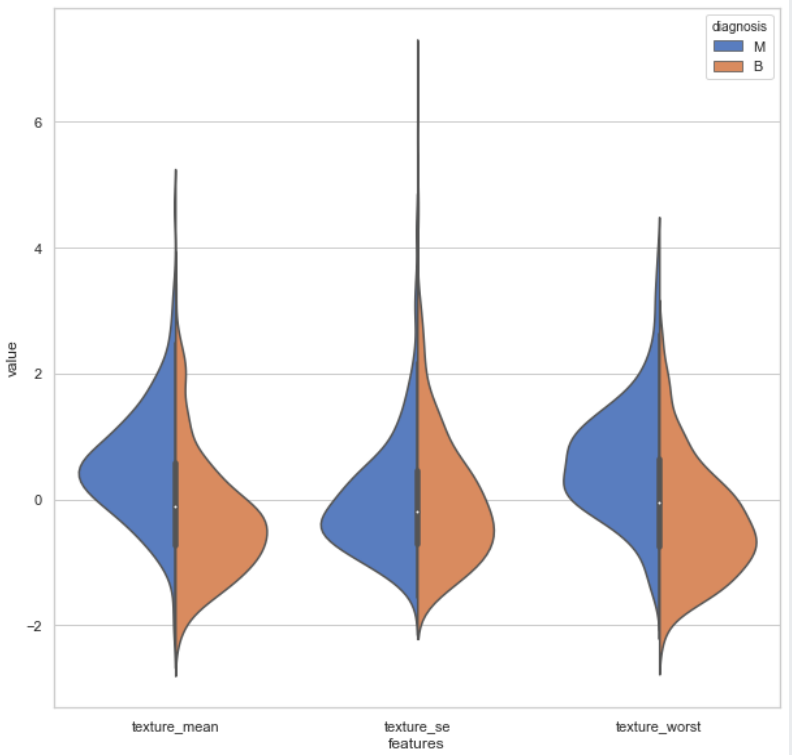
# Data at Glance

- The data set contains 32 columns and 569 data points. There are no missing values (null) in any of the features
- In the univariate analysis, it is shown that there is no categorical variable, and 30 continuous variables in the data set

# Data at Glance Continued

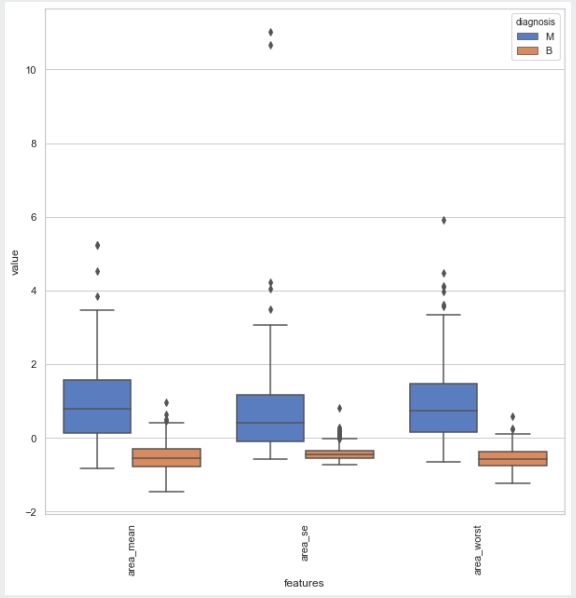
- It shows that the two-diagnosis distributed quite equally in both cases, indicating that within these feature, it is difficult to tell whether the tumor is benign or malignant just by looking at the data

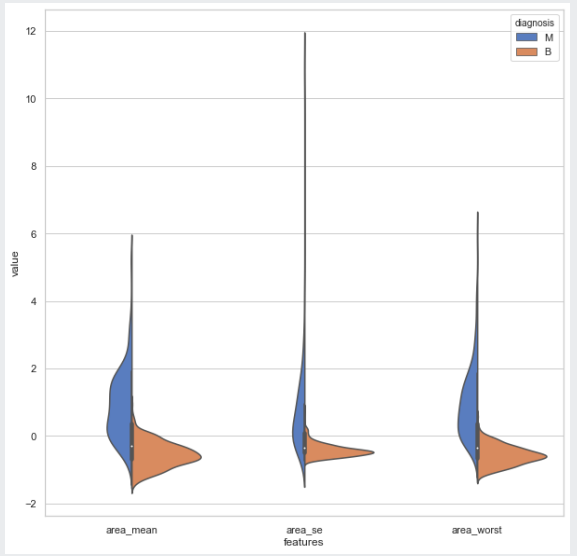


# Data at Glance Continued

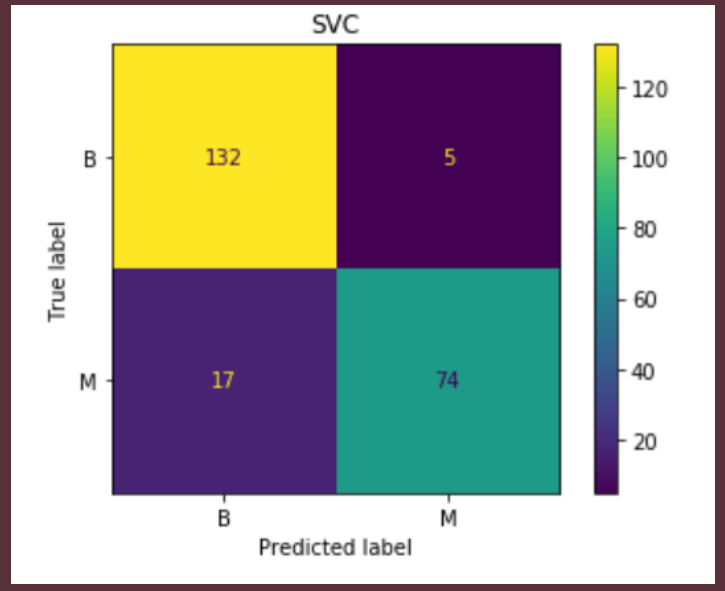
- , the figures below show a completely different story. It is easier to tell the two diagnosis just from looking at the data as the distributions of the two different groups are in different ranges. They are likely be selected as the features for the model

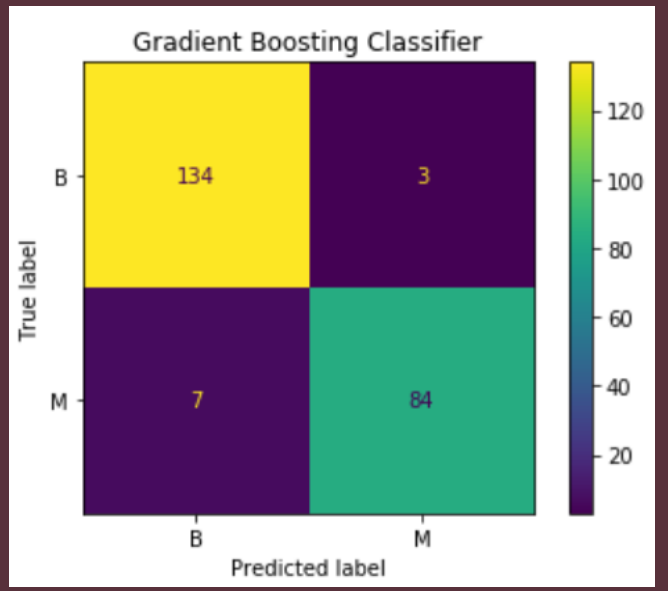


# Model Comparison

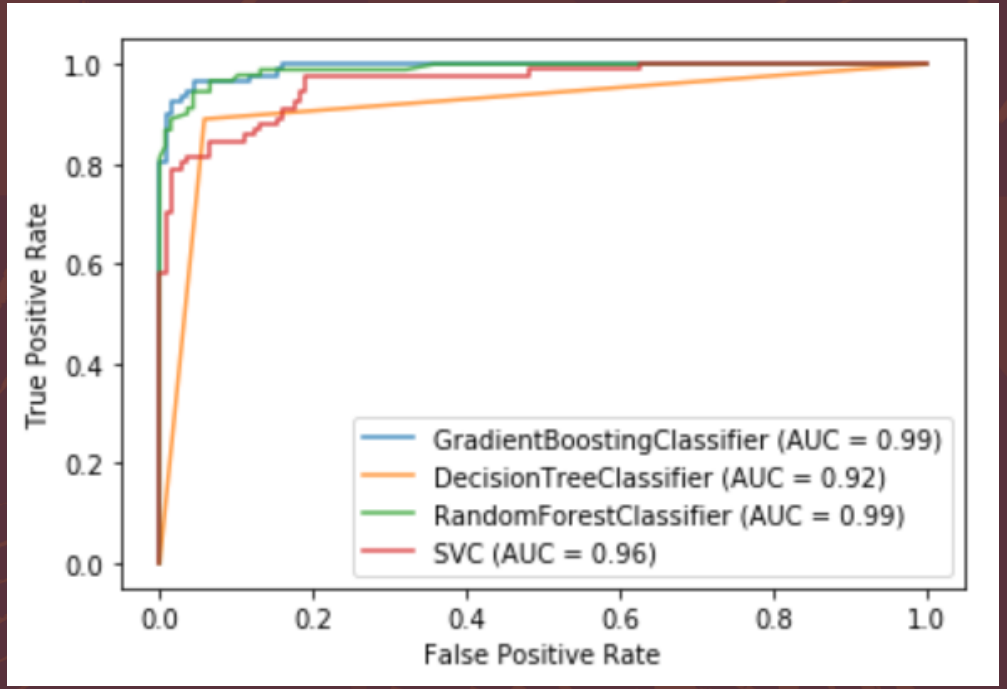
| Models | Accuracy Score |
|---|---|
| Support Vector Classifier | 0.903508 |
| Decision Tree Classifier | 0.921059 |
| Random Forest Classifier | 0.951798 |
| Gradient Boosting Classifier | 0.956140 |



# Model Comparison

- it is showing that both random forest classifier and gradient boosting classifier are better classifiers for this problem, while decision tree and SVC are not quite much.



# Summary and Next Step

- it is shown that both **gradient boosting classifier and random forest classifier** can perform good prediction for breast cancer data set. In the machine learning algorithm could be realized and implemented in the real-life situation could greatly benefitting the current healthcare system in treating breast cancer more responsively and accurately.
- There are some rooms for improvement in this project, such as tuning the hyperparameters for the gradient boosting or random forest classifiers to achieve even better result.