Capstone 2 Milestone Report: Breast Cancer in Wisconsin

Tzu-Ying Chen

Problem

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.

Goal and Utility

The goal of this project is to analyze the symptoms and tumors characteristics, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension that breast cancer could have at different stage, and provide prediction on whether the tumor is malignant or benign. The analysis and the prediction can quickly help doctor prescribe necessary treatment and healthcare.

Data

The data set for this project contains 569 observations including patient ID number, and the features of the cell nucleus, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

The data set is available on UCI Machine Learning Repository. The page contains information about the data set and its source and relevant scientific publications:

http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Approach

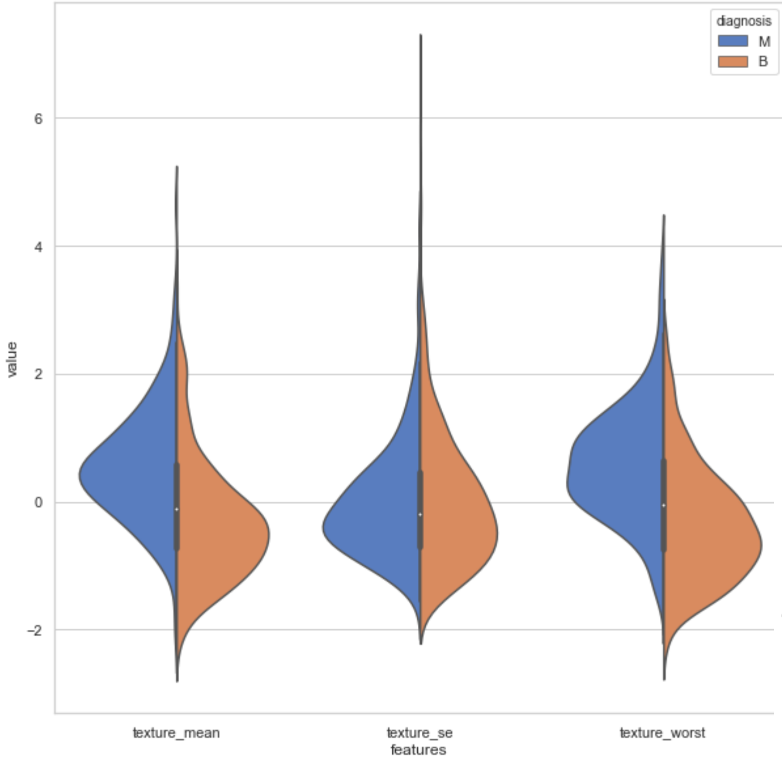
It is a supervised classification task. The diagnosis (whether the tumor is malignant or benign) will be our target variable, while the characteristics of the cell nucleus will be used as the predictive variables. Logistic Regression, K-Nearest Neighbor, Support Vector Machine, and Random Forest Classification are some of the machine learning models that could potentially be used in this project. Data exploratory analysis will provide deeper understandings of the data and address the problem more clearly, and cross-validation could ensure the outcomes of the model is reliable.

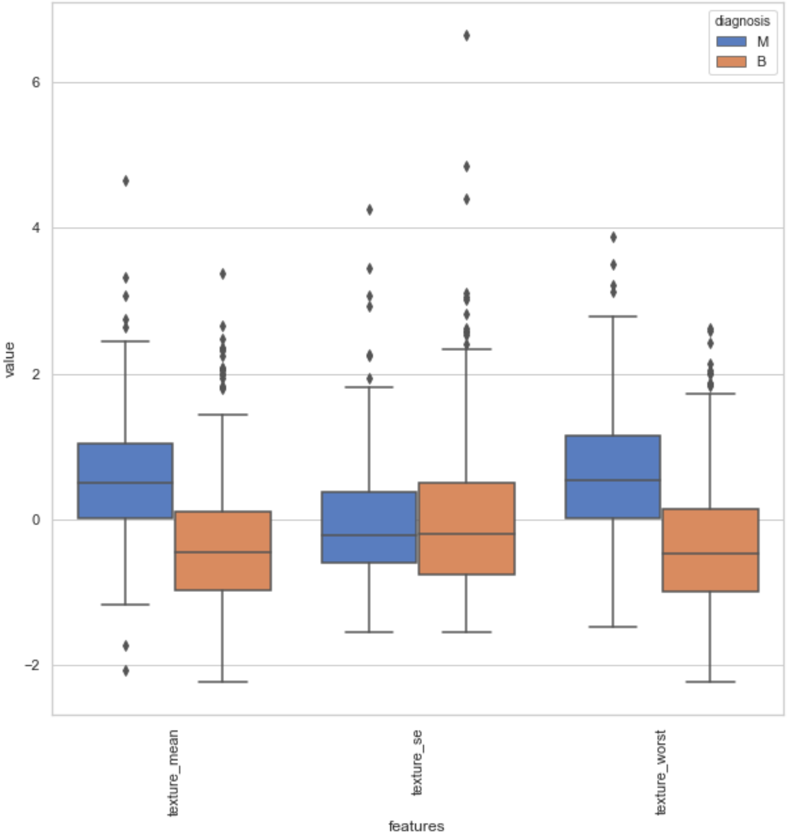
Deliverables

The deliverables for the project will include the source code, the data set, and a paper addressing the purpose, approach, findings and results of the project.

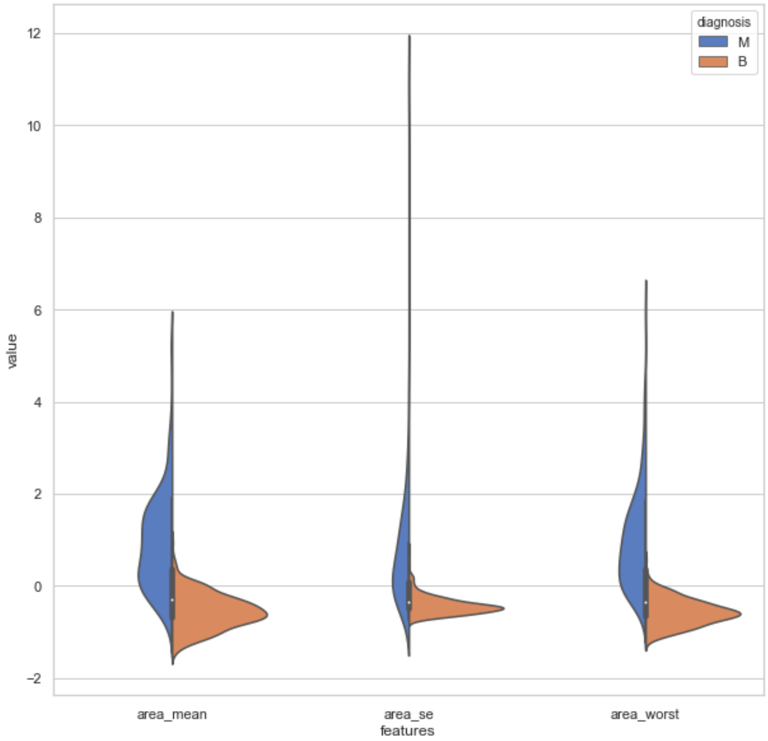
Exploratory Data Analysis

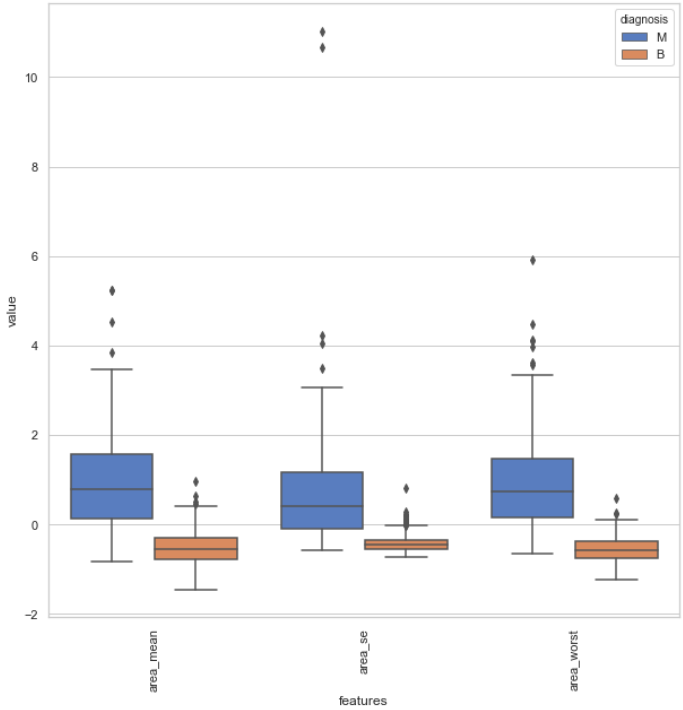
The data set contains 32 columns and 569 data points. There are no missing values (null) in any of the features, hence, it is not required to do much data imputation. In the univariate analysis, it is shown that there is no categorical variable, and 30 continuous variables in the data set. I excluded the feature “id” because the characteristics of id does not imply much information to the algorithm. Next, I am interested in knowing the distribution of each feature and identify if there is any outliers within the feature. The violin plots and box plots come in handy as it can quickly tell the distribution and outliers by feature and by the diagnosis (the target variable).



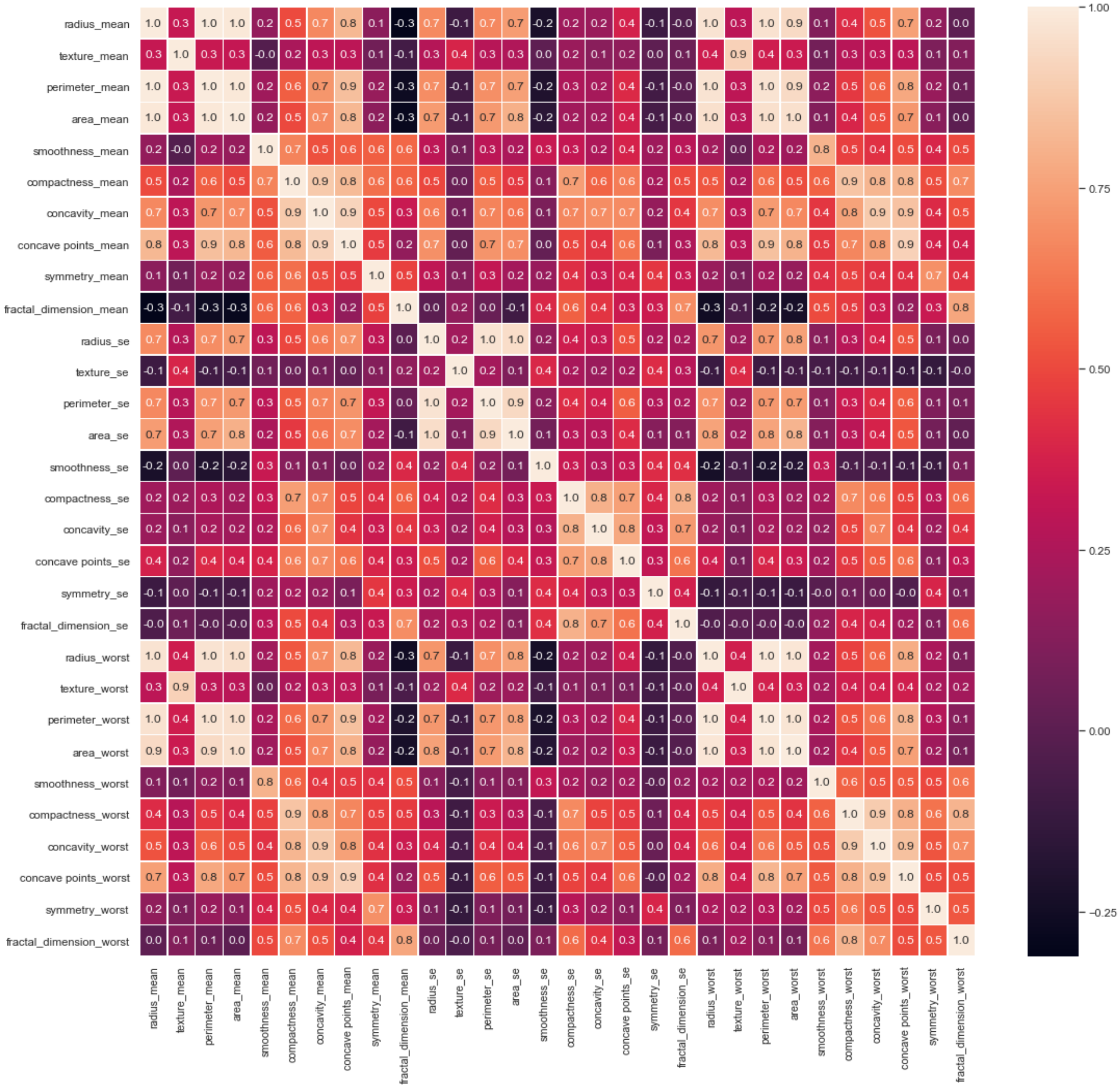


The above figures show the distributions of texture_mean, texture_se, and texture_worst features after normalization. It shows that the two-diagnosis distributed quite equally in both cases, indicating that within these feature, it is difficult to tell whether the tumor is benign or malignant just by looking at the data. It also indicates that in the feature selection, we could possibly just include one of these features if they are highly correlated to each other. On the other hand, 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.





Next, in the bivariate analysis, we can look at the correlation between each feature. The features that are more highly correlated could be “combined”, and just consider one of them to be the feature in the model. In this case, we can see that radius_mean with concave points_mean, area_worst, radius_se, perimeter_se, area_se, and concave points_worst. The same ideas apply to all other features. Hence, in this stage, we can come up with a set of features that will be input into the model, which are radius mean, texture mean, smoothness mean, compactness mean, concavity mean, symmetry mean, fractal_dimension_mean, radius se, texture se, smoothness se, compactness se, concavity se, concave points se, symmetry se, fractal dimension se, smoothness worst, compactness worst, concavity worst, symmetry worst, and fractal dimension worst.



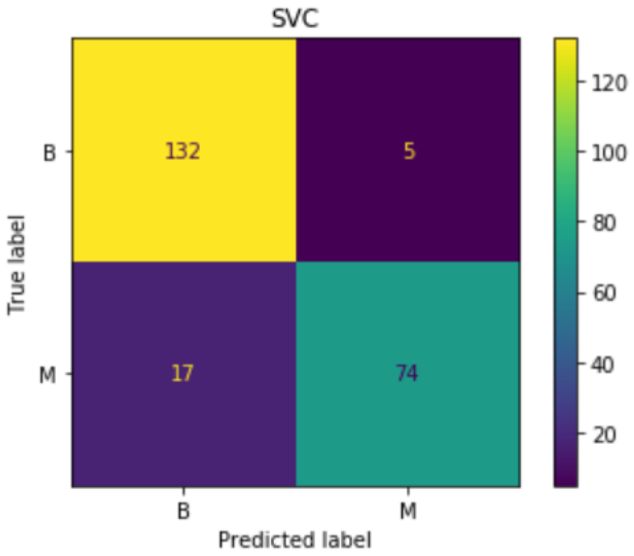
Machine Learning Models

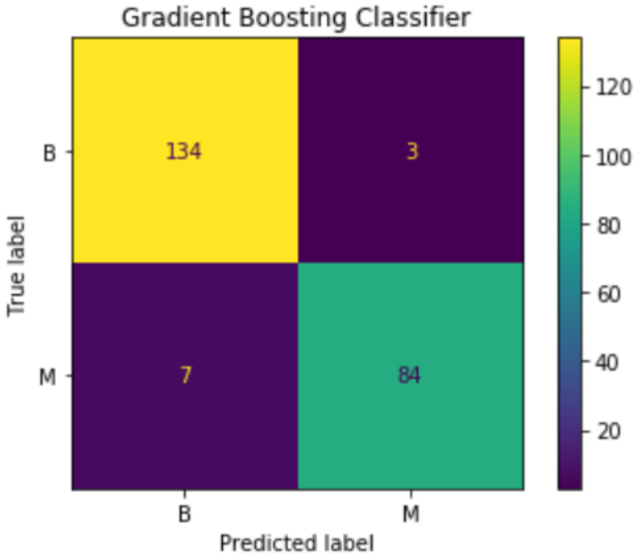
There are many way classification algorithms that could be apply to the problem, such as support vector classifier, decision tree classifier, gradient boosting classifier, and random forest classifier. In this project, I would like to explore different classifications algorithm and find the best model that can predict the breast cancer problem well.

To compare the results for the classification models, accuracy score is one of the metrics that help determines which model performs better.

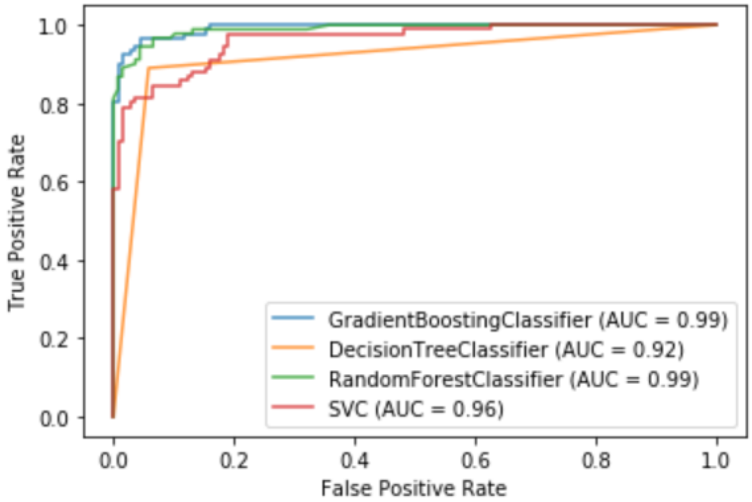
Models	Accuracy Score
Support Vector Classifier	0.903508
Decision Tree Classifier	0.921059
Random Forest Classifier	0.951798
Gradient Boosting Classifier	0.956140

From the above table, gradient boosting has the highest accuracy score while support vector classifier has the lowest. By looking at the confusion metrics, the support vector classifier does have higher false negatives (the type II error). While on the other hand, gradient boosting classifier does very well in terms of accuracy. Both the false positives and false negatives are low.





Another good metrics for classification model comparisons is the ROC curve metrics, which is commonly used in the binary classification problem. In the figure below, 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

After comparing four classification models results, 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. The implementation of the model could also potentially reduce the healthcare cost and wrong diagnosis, helping doctors make quicker diagnosis and apply necessary treatments. 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. The similar practice can also apply to other disease to help faster and more accurate diagnosis.