[DATE]

DISCOVER THE RIGHT MODEL FOR BREAST CANCER DIAGNOSIS RESULT

CAPSTONE PROJECT 1

YOHEITA YOSHIMURA

10/22/2021

Table of Contents

Introduction

Using the breast cancer diagnosis conducted at The University of Wisconsin, the target is to find the most accurate algorithm. The data was obtained after diagnosing patients with a fine needle aspirate (FNA) of a breast mass to produce a digitalized image to compute features.

A high accurate model may be used mainly for doctors to help finalize the result of the patient's digital image before taking a cells sample decreasing physical damage to the patent.

The data of fine-needle aspiration can be found from:

LINK: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

I have used Python for this project. Implemented detail can be found in the below link:

LINK:

https://github.com/yoheita/Springboard/blob/add2a90c6db47963a269323004f56c98d2671fac/Capstone%20Project%201/Actual_Capstone_Project_1_mod-Copy1.ipynb

Data Acquisition and Wrangling

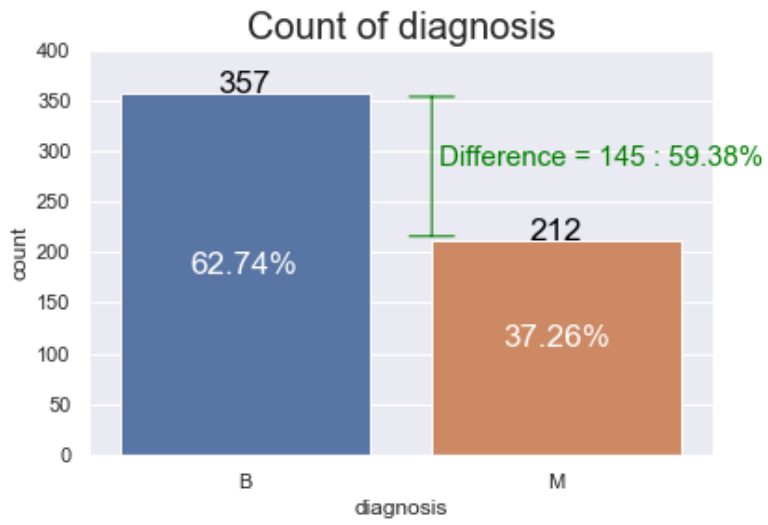
To generally understand the dataset, I have converted data into panda DataFrame. Dataset with panda DataFrame, enable me to understand the general shape of the dataset and to view a sample of patient's data, provided columns, and column's datatype.

Dataset was also wrangled by cleaning the dataset. Data cleaning was conducted to remove the unnecessary features and check whether null data exists in the dataset. Checking for null data and removing unnecessary features is an important task before analyzing data.

Storytelling and Inferential Statistics1

With purified data, I had to understand further with the dataset.

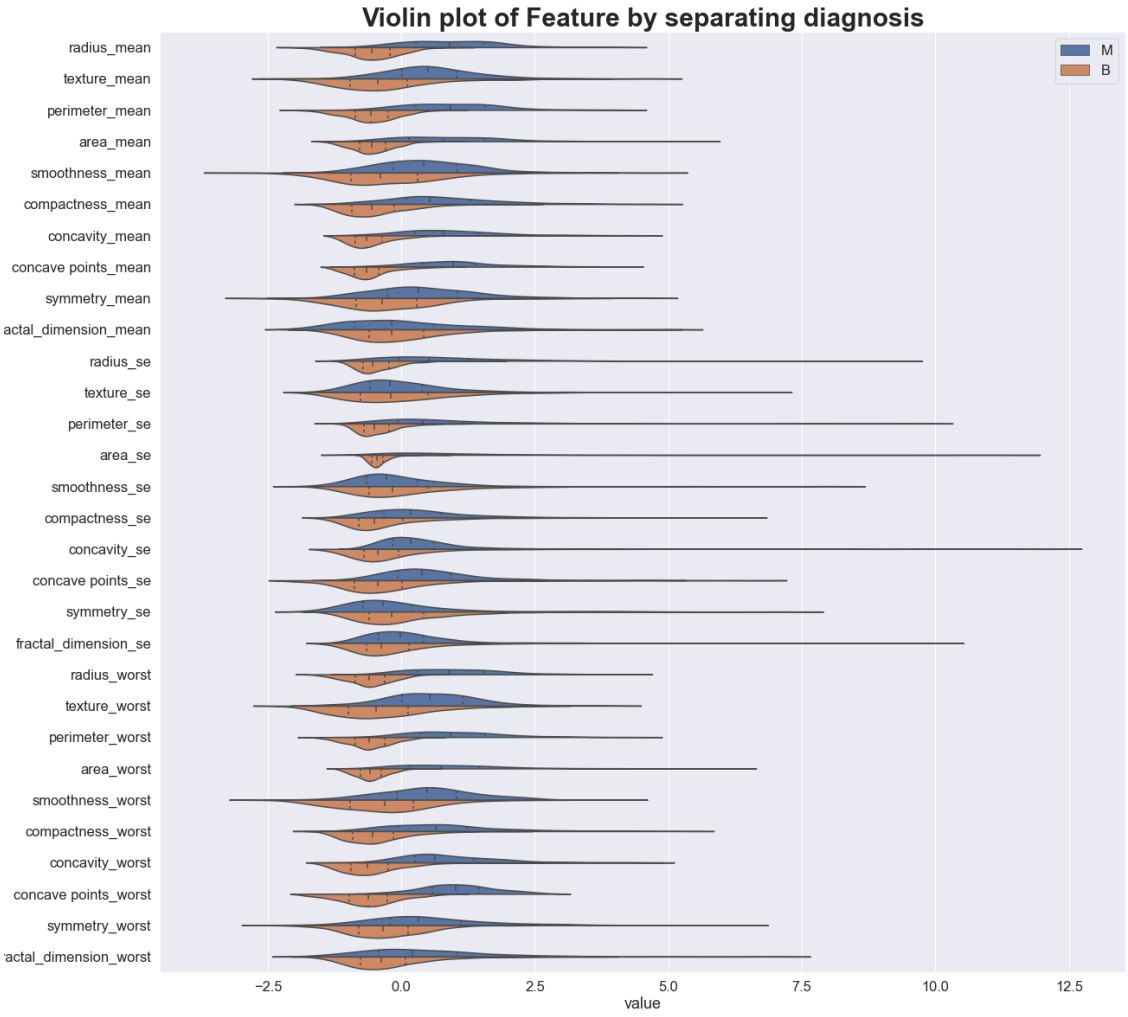
The target data is a feature called 'diagnosis', and data was displayed either 'M' as Malignant or 'B' as Benign. The below graph shows the difference in the count with the 'diagnosis' target feature.

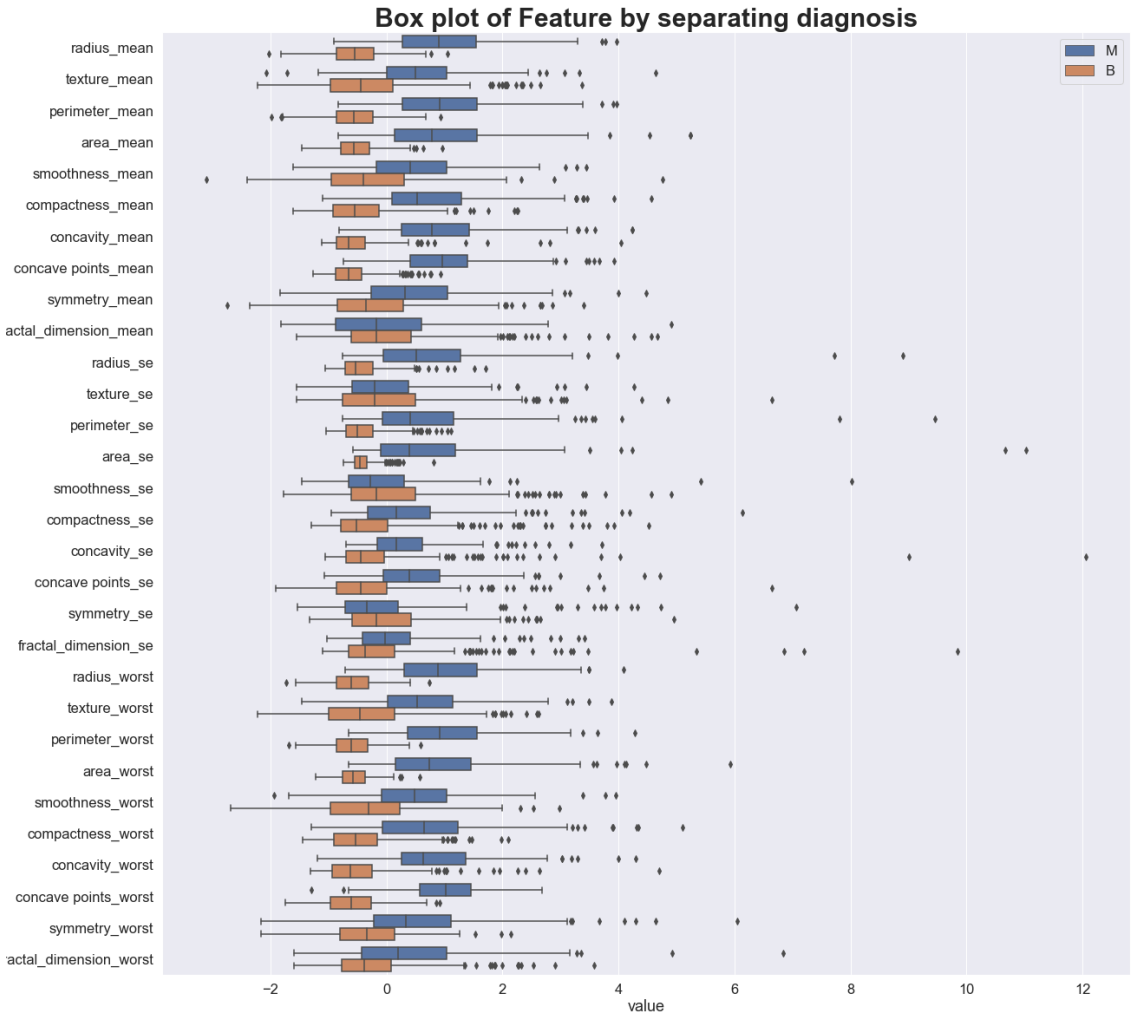


The graph represents the target variables containing an unbalanced number of Malignant and Benign. There are 212 patients (37.2%) diagnosed as Malignant and 57 patients (62.74%) were

diagnosed as Benign. By comparing to total, there is a 59.3% difference between Malignant and Benign.

For checking the distribution of each feature with separating diagnosis results, I used standardization to compare each feature. Standardization enables plotting a graph of violin plot for distribution comparison and box-plot for visualizing the distribution of quantitative values by displaying median, Q1, Q3, maximum, and minimum.

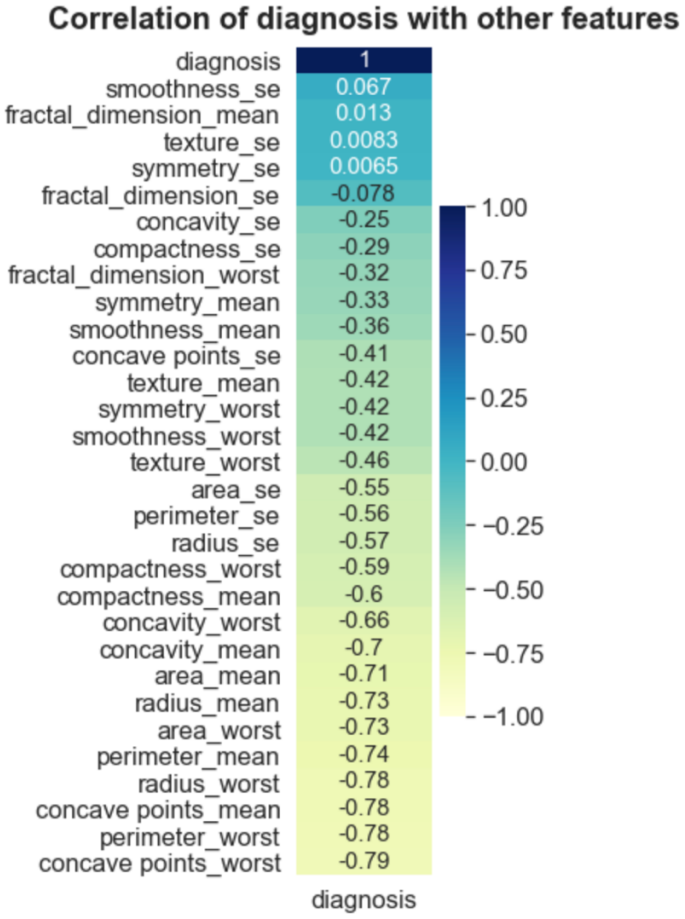

Violin plot of Feature by separating diagnosis
M
B
radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave points_mean
symmetry_mean
actal_dimension_mean
radius_se
texture_se
perimeter_se
area_se
smoothness_se
compactness_se
concavity_se
concave points_se
symmetry_se
fractal_dimension_se
radius_worst
texture_worst
perimeter_worst
area_worst
smoothness_worst
compactness_worst
concavity_worst
concave points_worst
symmetry_worst
actal_dimension_worst
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
value



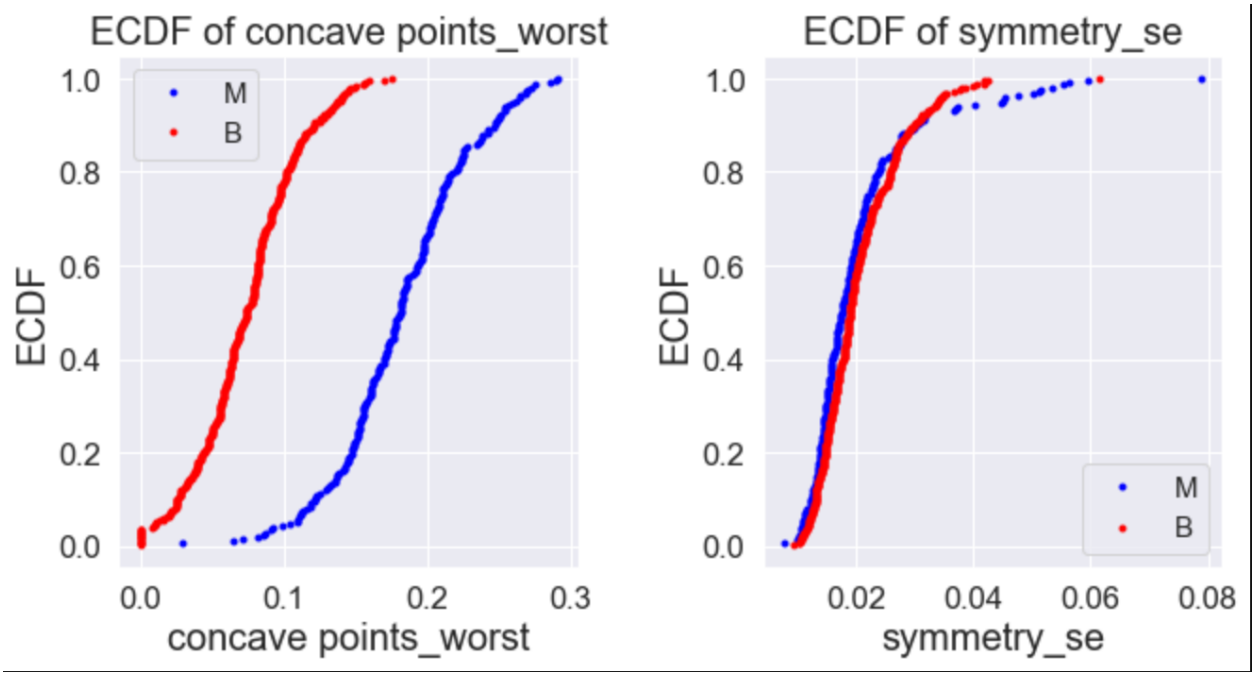
The graph presents some of the features that have similar distribution such as "smoothness_se" and "symmetry_se". Both features have a similar distribution pattern of Malignant and Benign which could mean low correlation to the target variable.

Whereas, Malignant and Benign of "concave points_worst" has a similar pattern of distribution with different mean. Since Malignant and Benign means are separated, the correlation might be high.

Heat map enables to display of the correlation of each feature to the target feature ”diagnosis”.



The highest negative correlation was “concave points_worst" with a negative correlation of -0.79, and the lowest correlation of feature is "symmetry_se" with a positive correlation of 0.0065. This describes the relationship between the distribution of the diagnosis and correlation. I have compared the highest and lowest correlation with the empirical cumulative distribution of features (ECDF).



We can see that 'symmetry_se' Malignant and Benign have a very similar distribution unable to distinguish diagnosis from this data.

Baseline Modeling

Since target data is categorical and has a binary classification, I applied Logistic Regression Algorithm. To apply the model, I have separated the dataset into test data and train data. Train data will be used to train a model based on the algorithm (in this case Logistic Regression) to apply to test data to create predicted data and actual testing data.

To separate the data, 70% were used for Train data and the rest of 30% was used for Test data. As I recall the ratio of M and B is not balanced. Therefore, I made sure to separate with the same ratio to produce the closest environment to overall data.

After applying Logistic Regression, the result was the following:

I have outputted the classification report which outputs the basic statistic result of the model that I have used.

The table's term is follow:

precision: It is the ratio of true values and a total number of predicted true values. True value is the outcome of predicted data is equal to the target data (diagnosis in this case). The equation is:

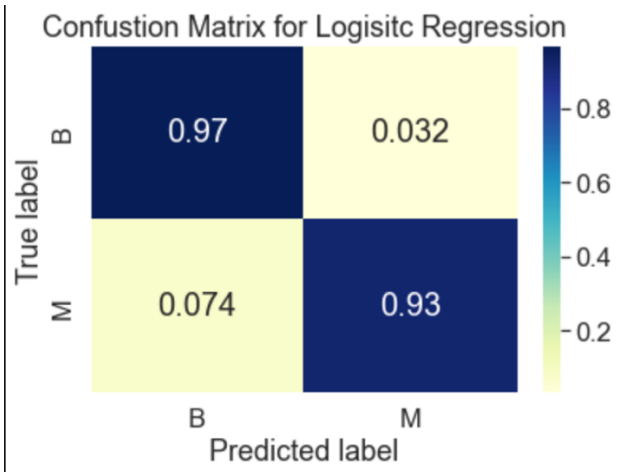
Recall: it is the ratio of true values and an actual number of true in the data.

F1 score: It is a combination of precision and recall, it presents the overall score of the model.

Support: it is the actual M and B data from the test

Logistic Regression

	precision	recall	f1-score	support
B	0.94	0.98	0.96	107
M	0.97	0.89	0.93	64
accuracy			0.95	171
macro avg	0.95	0.94	0.94	171
weighted avg	0.95	0.95	0.95	171

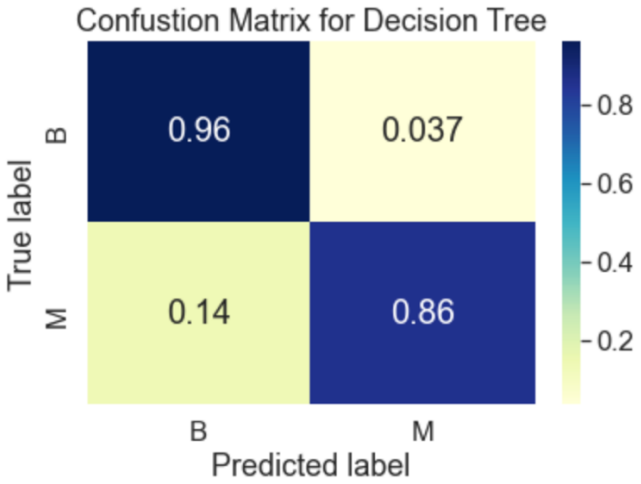


Extended Modeling

Since target data is categorical, I also applied Decision Tree, SVM, Random Forest using the same test and train data from Logistic Regression.

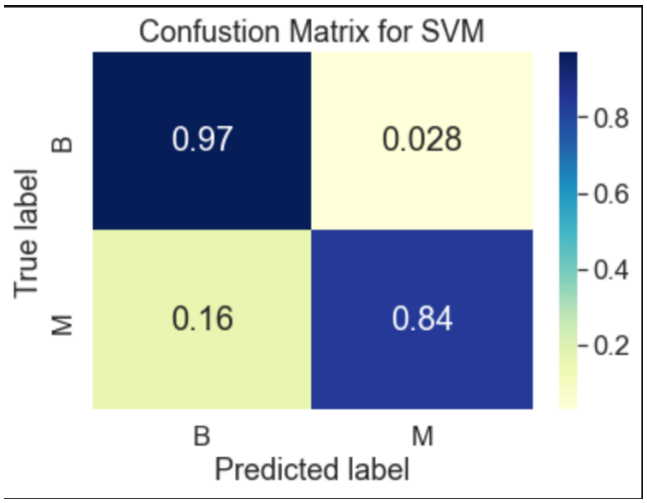
Decision Tree

	precision	recall	f1-score	support
B	0.93	0.96	0.94	107
M	0.93	0.88	0.90	64
accuracy			0.93	171
macro avg	0.93	0.92	0.92	171
weighted avg	0.93	0.93	0.93	171



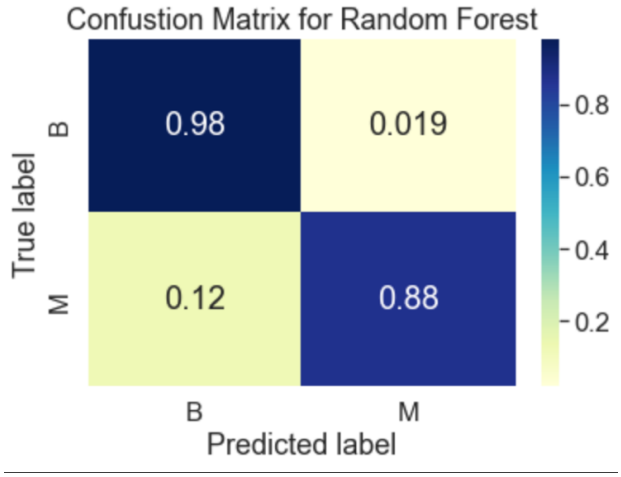
SVM

	precision	recall	f1-score	support
B	0.88	0.99	0.93	107
M	0.98	0.78	0.87	64
accuracy			0.91	171
macro avg	0.93	0.89	0.90	171
weighted avg	0.92	0.91	0.91	171



Random Forest

	precision	recall	f1-score	support
B	0.93	0.98	0.95	107
M	0.97	0.88	0.92	64
accuracy			0.94	171
macro avg	0.95	0.93	0.94	171
weighted avg	0.94	0.94	0.94	171

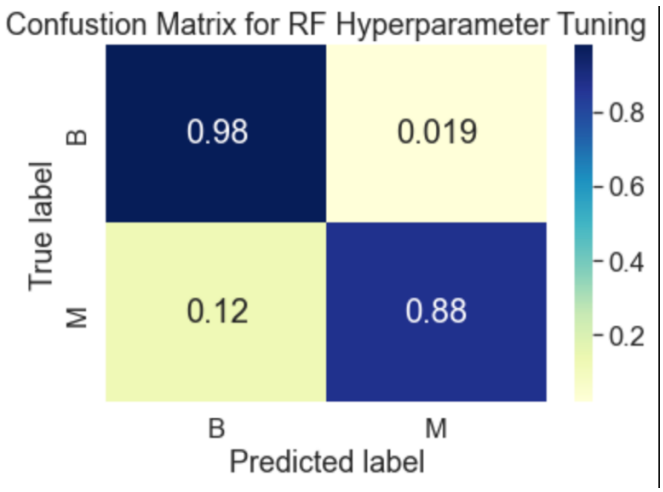


Hyperparameter Tunning for Random Forest

Conducted Hyperparameter Tunning to Random Forest to search for better parameters.

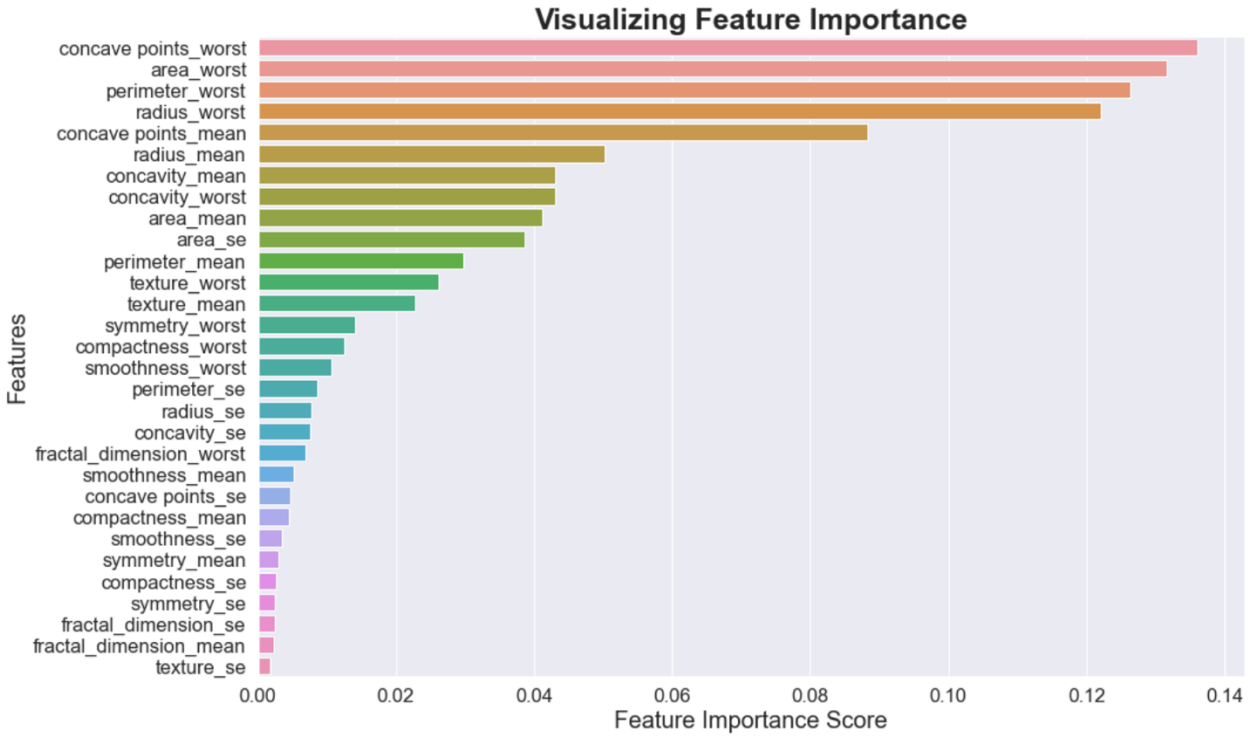
The result became the following:

	precision	recall	f1-score	support
B	0.93	0.98	0.95	107
M	0.97	0.88	0.92	64
accuracy			0.94	171
macro avg	0.95	0.93	0.94	171
weighted avg	0.94	0.94	0.94	171



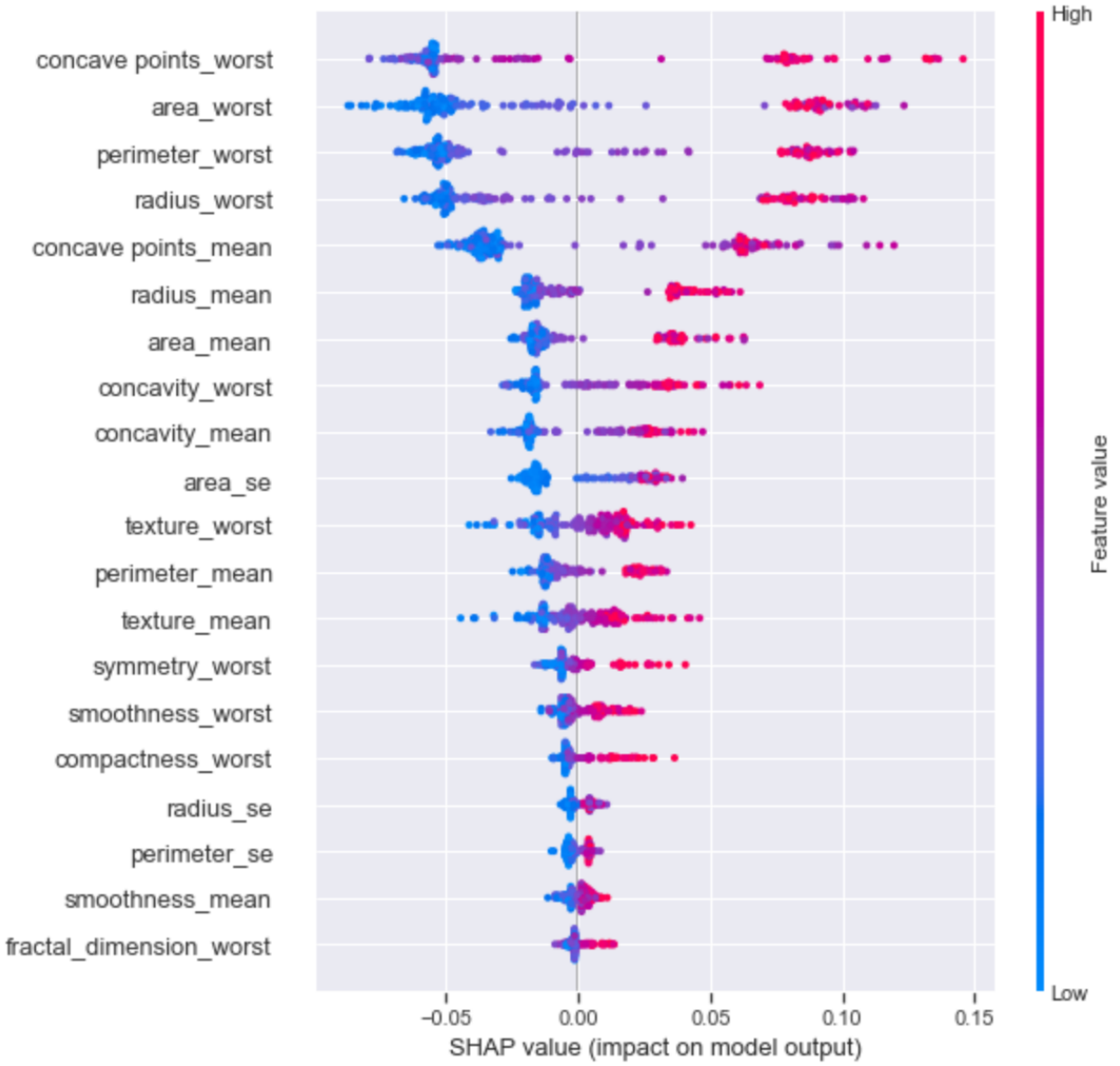
Findings

By using Random Forest from sklearn, the model enabled to present a list of important features which impacted the result of diagnosis from the model.

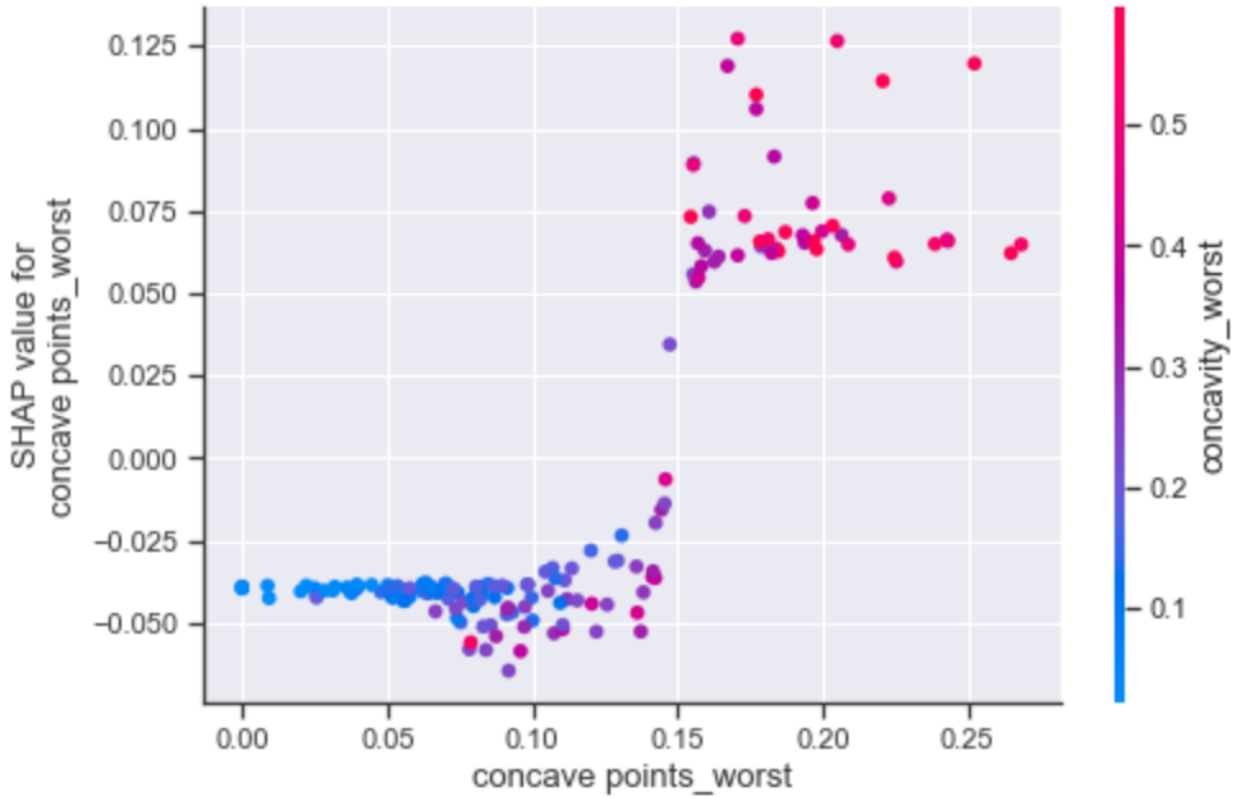


We can see that " concave points_worst " has the highest feature importance.

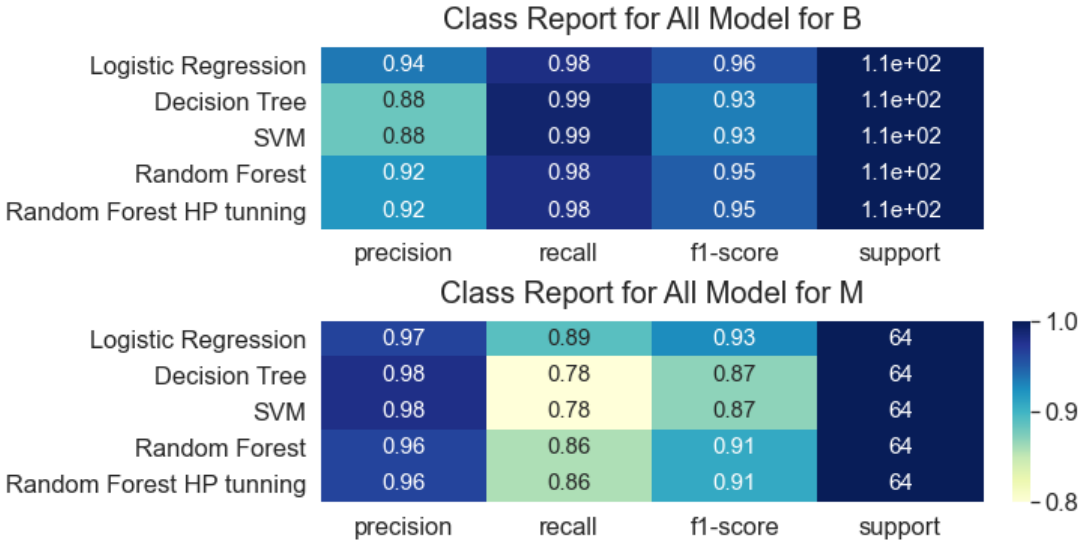
I imported SHAP to present how each feature was impacting diagnosis results ‘M’ and ‘B’. SHAP summary_plot generates how each feature impacts the predicted result. The color bar indicates the feature importance level. Each dot represents each data, and the color of the dot towards red means an important feature for predicting Malignant and Benign. The X-axis represents the level of impact on the model using the SHAP value.



Since “concave points_worst” has the highest feature importance, output the single feature against the SHAP value for understanding the importance of each data point and the distribution of the graph.



Now I will compare each classification report to the table to understand which algorithm had the highest accuracy score.



We can see that Logistic Regression Algorithm has the highest scores (precision, recall, and f1 score) for Benign and Malignant with an f1 score of 0.96 with 0.93.

Conclusions and Future Work

Upon investigation, I managed to find Logistic Regression has the highest accuracy score with an f1-score over 0.9. F1-score over 0.9 defines above 90% of the prediction is the same result as the current dataset. Also, I managed to discover the most impacted feature “concave points_worst” for prediction using SHAP and Random Forest. This project can conclude that the highest algorithm model is Logistic Regression and the most important feature for prediction is “concave points_worst”.

If more time was allowed, I would have investigated deeper for this project. Firstly, I would have used more algorithms to investigate for higher accuracy scores. Since there is more algorithm that can be applied for binary classification. Second, I would have used more SHAP graphs to understand the dataset deeper. Using the SHAP graph I created, there were few data with low feature value but high SHAP value or the opposite. Understanding the reason behind this result would give a deeper understanding of the feature. Therefore, with more time allowed, I would have tested more algorithms and SHAP for a deeper understanding of the data.

Recommendations for the Clients

- Able to assist to predict patient's condition before test result giving less stress.
- The model was able to predict over 90% with the current dataset.
- The highest impacted feature "concave points_worst"
    - This feature could be what doctors and patients should look out for the most.