Springboard—DSC
Capstone Project 1
BREAST CANCER PREDICTION

By: Yoheita Yoshimura
November 2021

Introduction

- Used a dataset from University of Wisconsin, which was produced from digitalized images of fine needle aspirate (FNA) with specimen cells characteristic and diagnosis results.
- FNA study shows:
  - False positive result: 0-2.5%
  - false negative result: 5-10% or even higher than 15%
- Explored classification models to predict diagnosis result using Python
- The project was developed under the classical data science method: data wrangling, exploratory data analysis (EDA), baseline modeling, and extended modeling

Data Acquisition and Wrangling

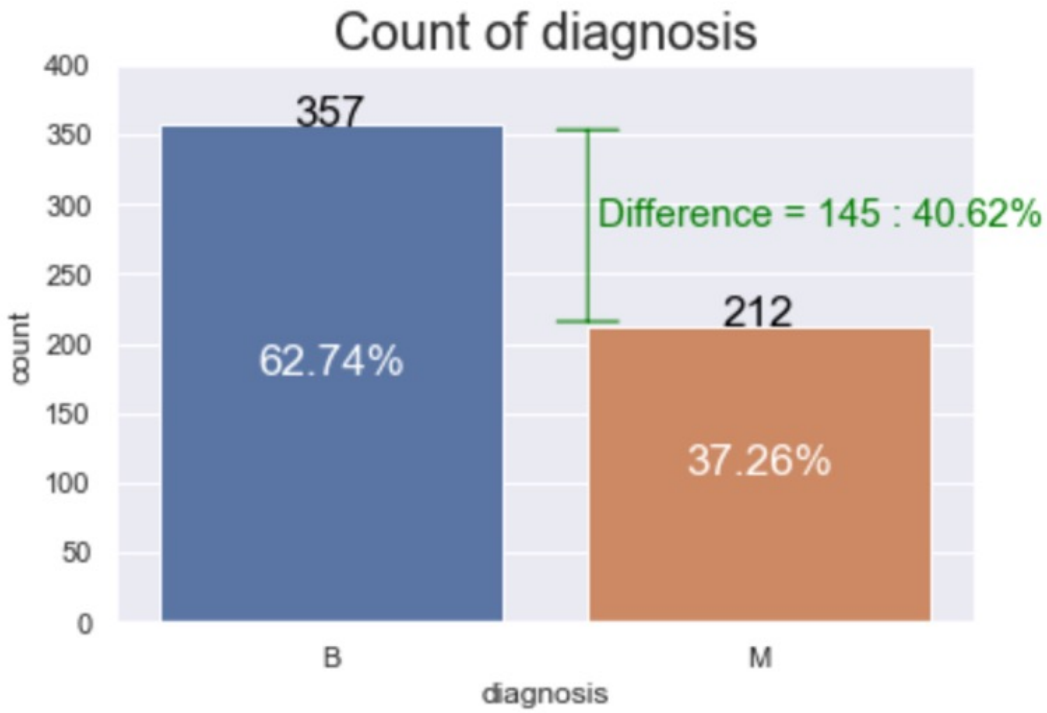
- 569 patient data and 13 features are included in the original dataset
- 13 features includes below data type:
  - 11 float data
  - 1 char data
  - 1 null data
- Target feature is called “diagnosis”
  - ‘M’ as Malignant or ‘B’ as Benign

Removed unnecessary features (ID and Unnamed: 32)

- Investigate for any null values in each feature

EDA: Storytelling and Inferential Statistics

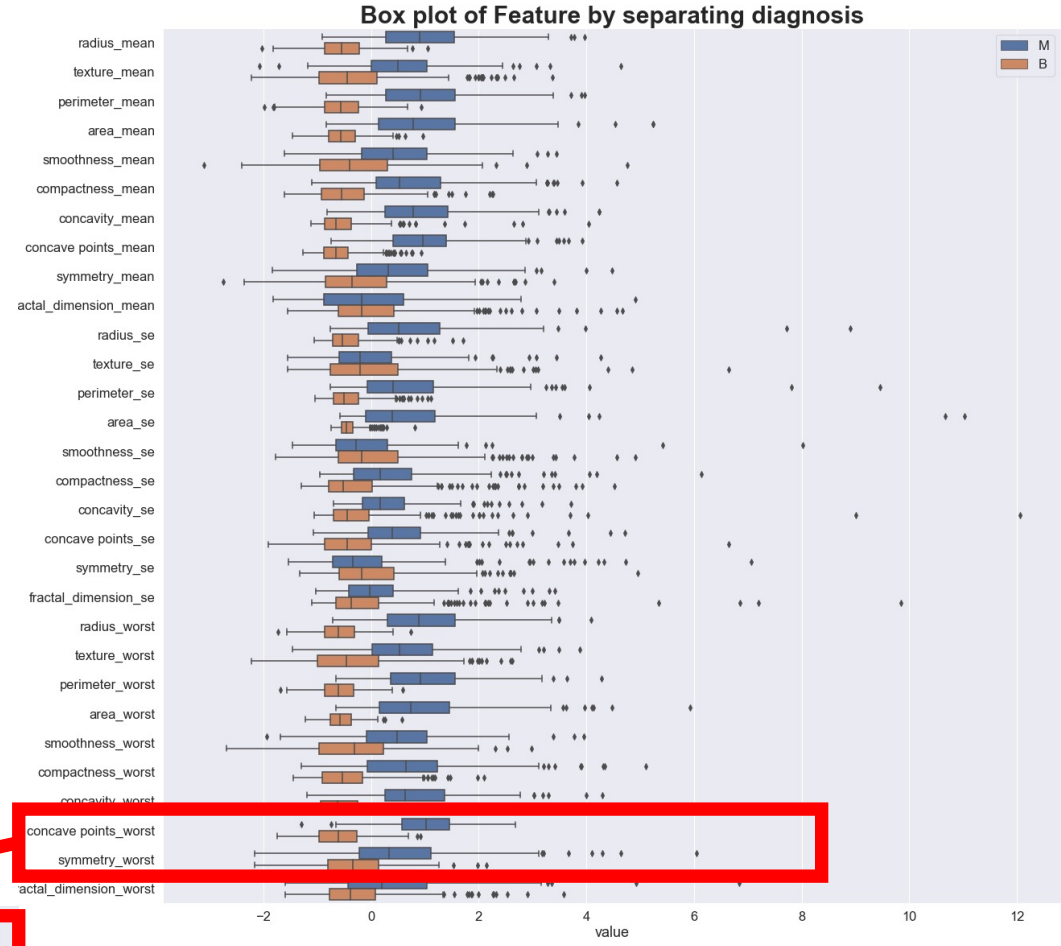
- Imbalanced data
  - Malignant = 212 patients (37.2%)
  - Benign = 357 patients (62.74%)



EDA: Storytelling and Inferential Statistics

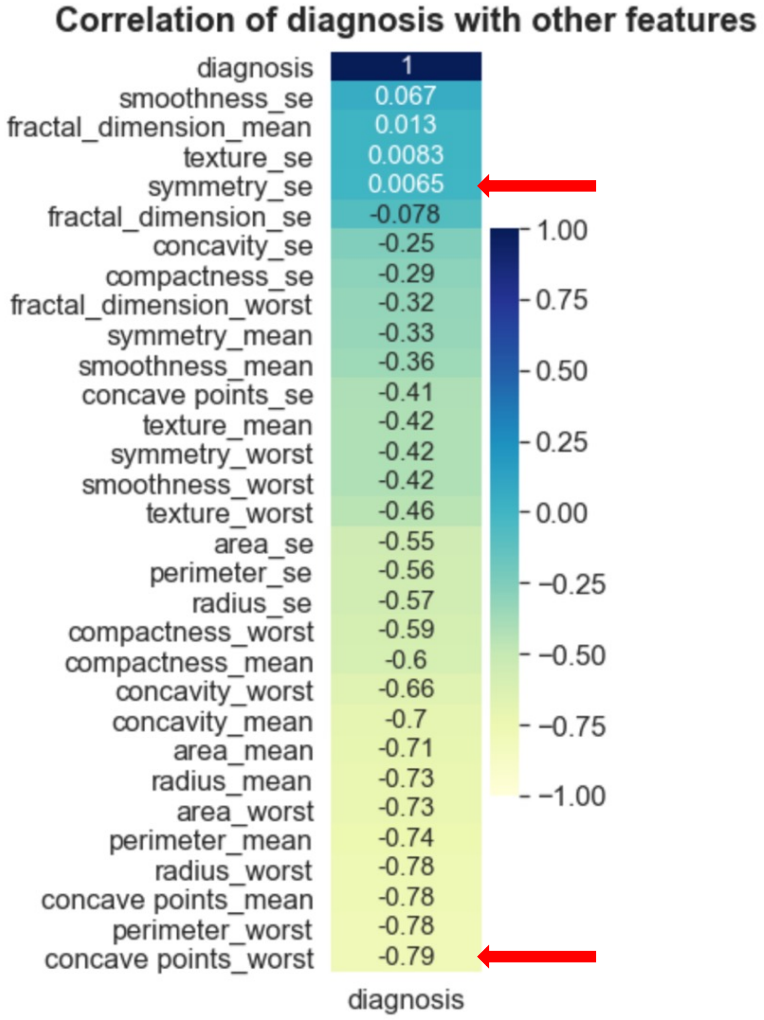
- Used standardization to compare each feature

Feature	Pattern of distribution	Mean	Predicting correlation
concave points_worst	similar	separate	High
symmetry_se	similar	similar	Low



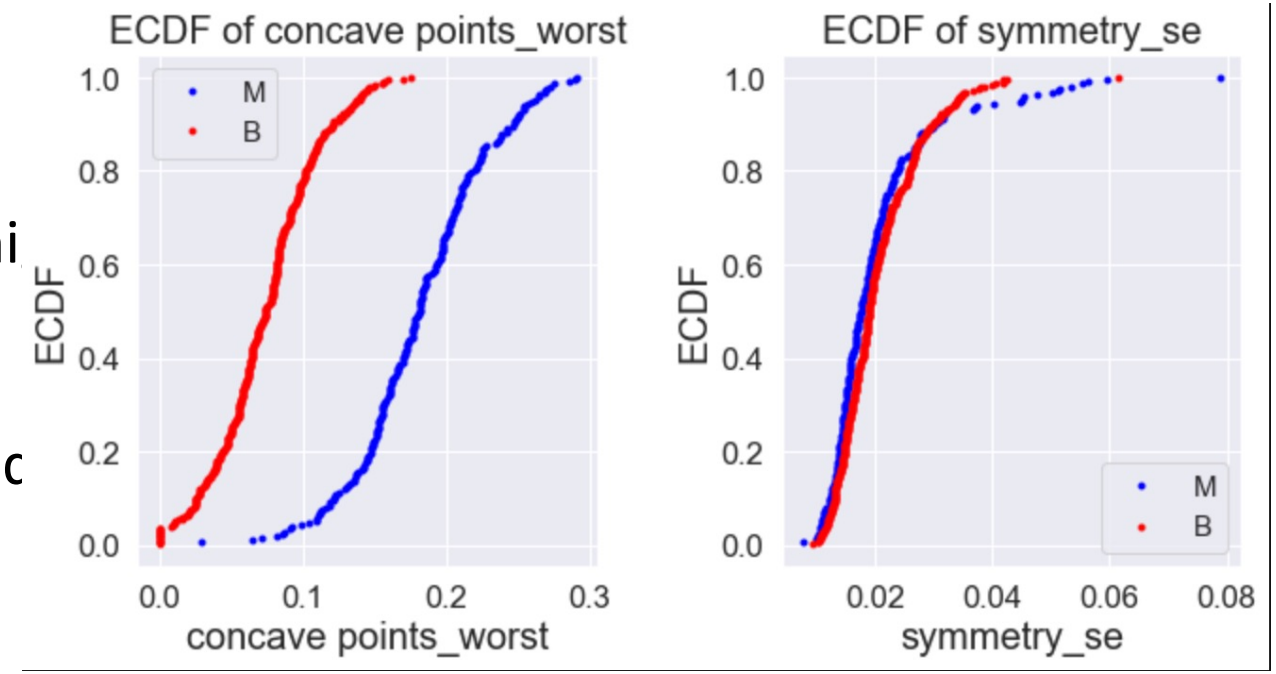
EDA: Storytelling and Inferential Statistics

- Plotted heat-map for correlation of target feature with other features
  - High negative correlation for concave point_worst (-0.79)
  - Close to zero symmetry_se (0.0065)



EDA: Storytelling and Inferential Statistics

- Plotted empirical cumulative distribution of features (ECDF) for:
  - “concave points_worst"
  - "symmetry_se"
- Able to also observe:
  - “symmetry_se” Malignant and Beni have a very similar distribution
  - “concave points_worst” separate distribution between Malignant anc Benign

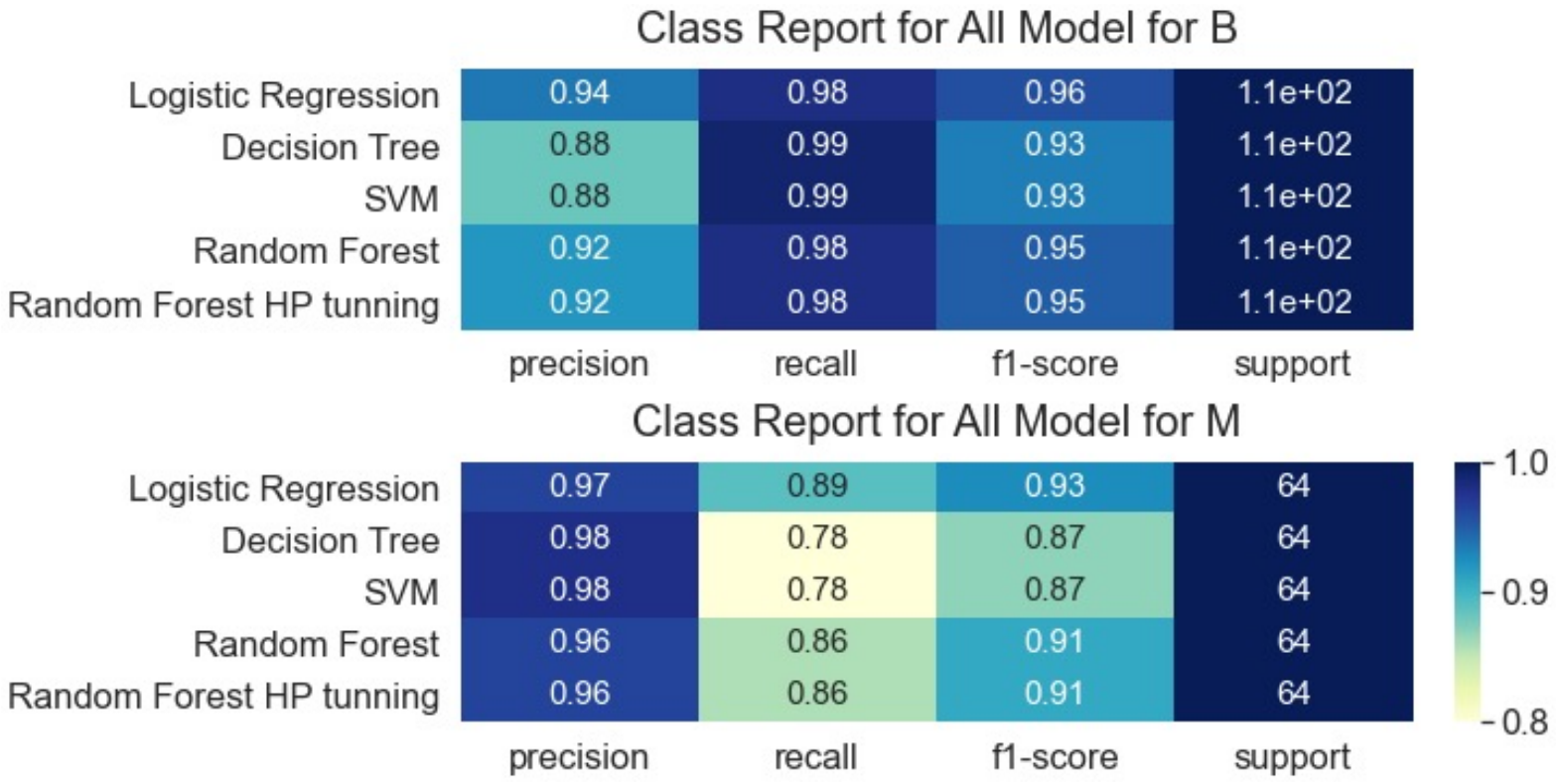


Modeling

- Separate the data:70% for Train data and 30% for Test data
- Keep ratio of Malignant or Benign in Train data and Test data is same as original data
- Used classification algorithms:
  - Logistic Regression
  - Decision Tree
  - SVM,
  - Random Forest
  - Random Forest (Hyperparameter tunning)

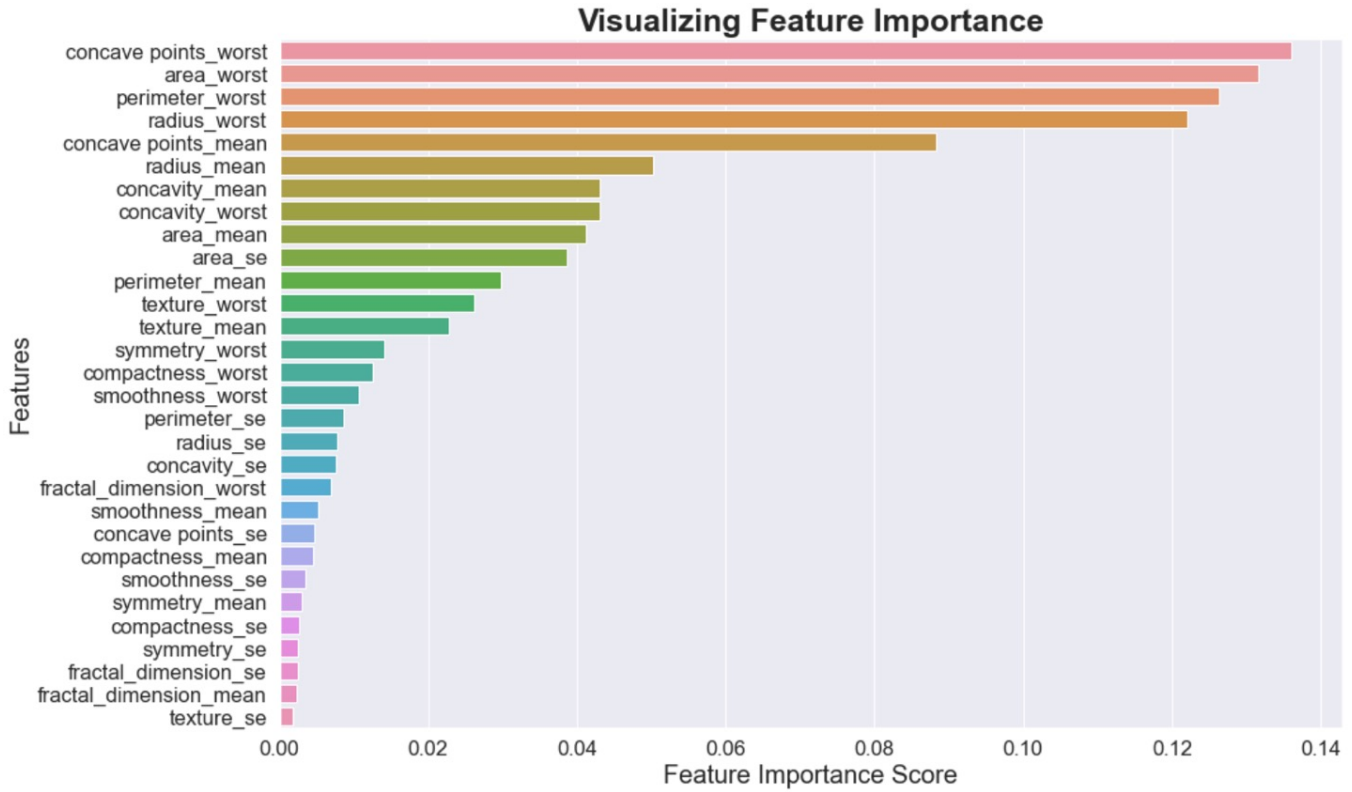
Modeling

- Logistic Regression scored highest accuracy for both Malignant or Benign



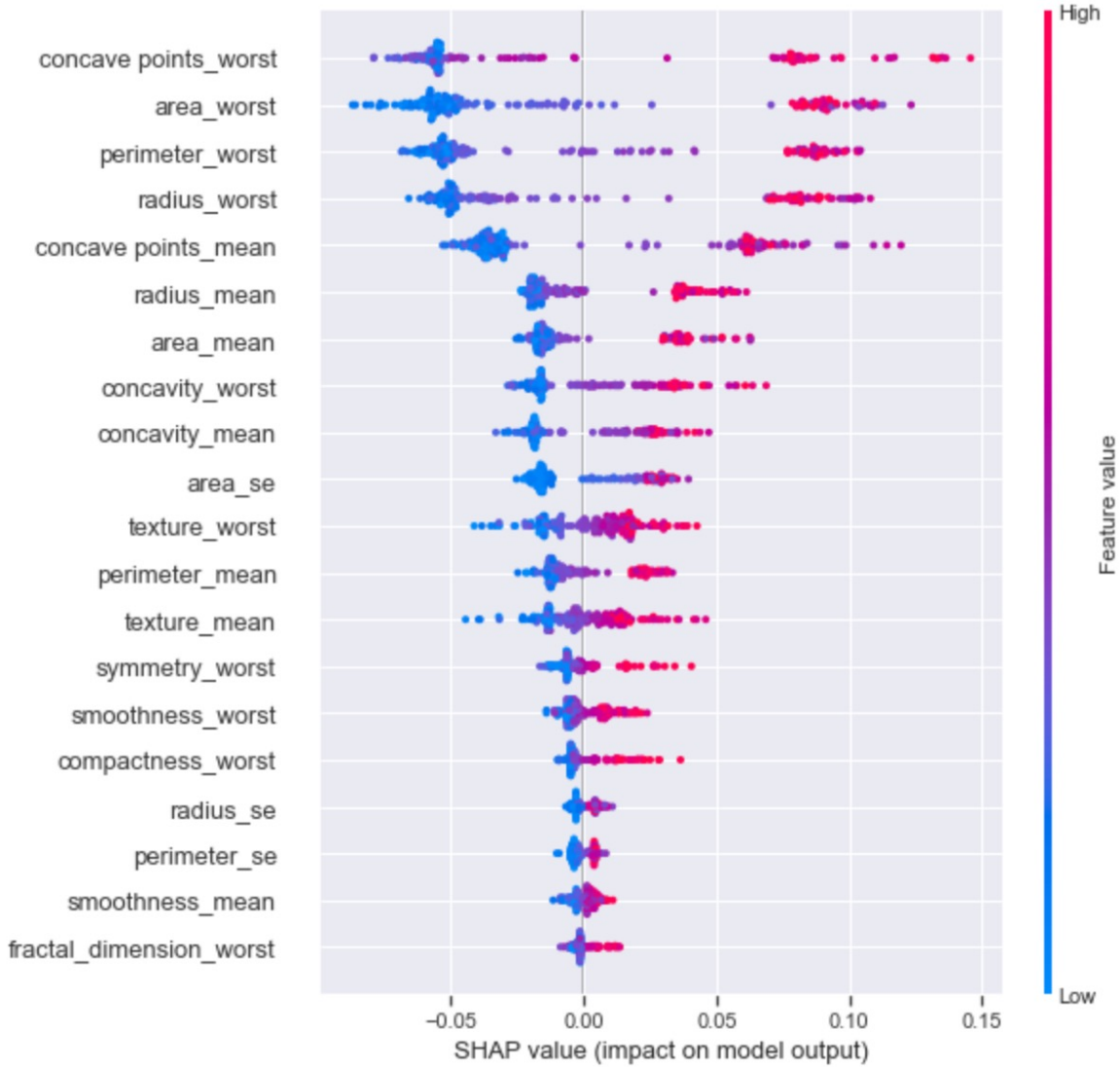
Findings

- Plotted feature importance using Random Forest
  - concave points_worst had highest feature importance



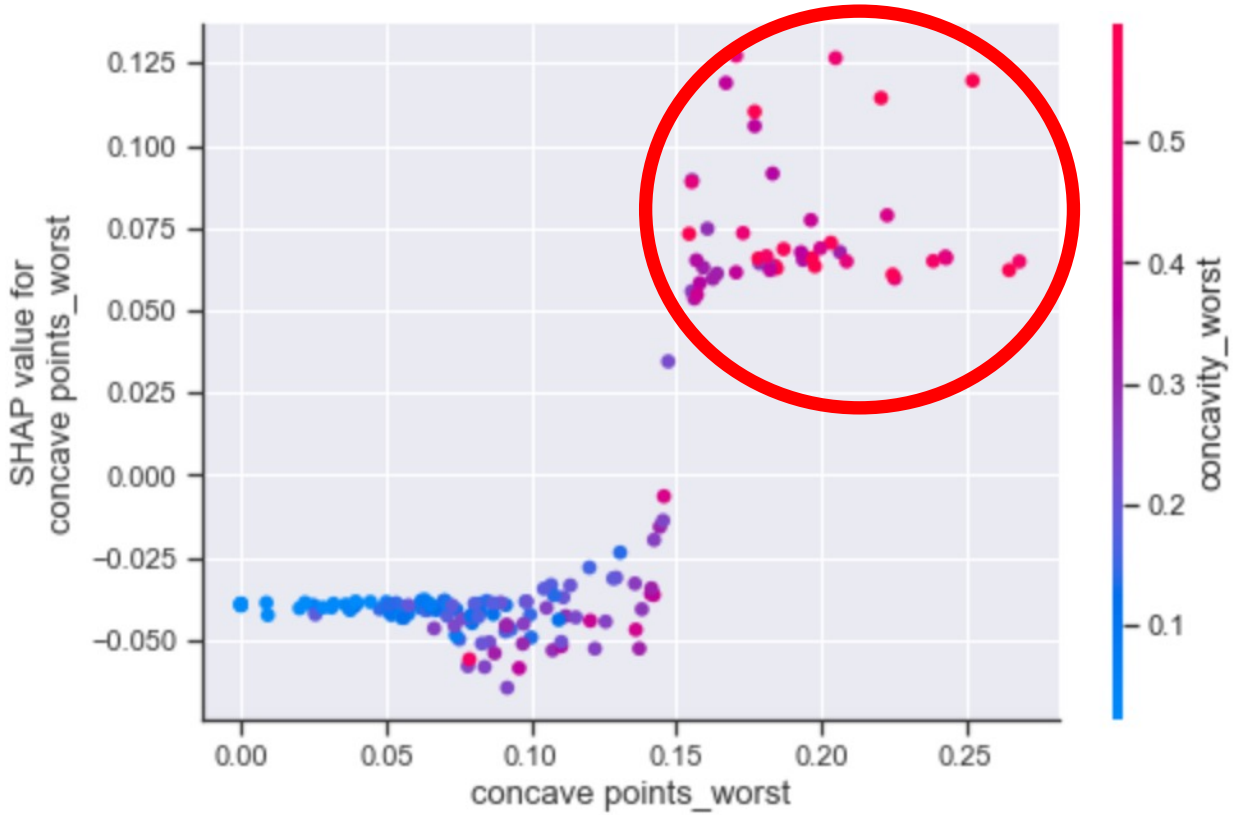
Findings

- Used SHAP to visualize feature importance
  - Importance is measured with SHAP value
- Discovered “concave point_worst” was still the highest importance



Findings

- Plotted feature value importance with “concave point_worst” with “concavity_worst” as color bar
- Both feature value likely to increase at:
  - “concave point_worst” > 0.15
  - Increasing the probability of tumor being malignant.



Conclusions - Overview

- Able to explore data with multiple classification algorithm
- Able to analyze the dataset and applied to the feature impact
- More in-depth study is necessary to use more effective algorithm
- Logistic Regression has the highest accuracy score:
  - f1-score over 0.96
  - False positive and false negative less than 1%

Recommendations for the Clients

- The model with Logistic Regression was able to predict over 90% with the current dataset.
- The highest impacted feature “concave points_worst” for diagnosis prediction
- If more data is provided, there is a possibility that a more accurate model could be made.