**Prostate Cancer Classification**

1. **Introduction**

Prostate cancer is the most commonly diagnosed cancer in Australia; it is estimated that more than 24,200 males were diagnosed in Australia with this form of cancer in 2022. If diagnosed early, prostate cancer has one of the highest survival rates. Growths in the prostate can be benign (not cancer) or malignant (cancer).

The task in this assignment is to investigate how well three different machine learning algorithms can predict whether the result of a biopsy is benign or malignant using a number of independent variables.

The dataset that we are going to use is a modified version of a small prostate cancer dataset that contains missing values. This data consists of 10 variables about prostate cancer tumors taken from 100 biopsies. These variables are:

* Id
* diagnosis\_result
* radius
* texture
* perimeter
* area
* smoothness
* compactness
* symmetry
* fractal\_dimension

The dependent variable 'diagnosis\_result' is a categorial one where 'B' stands for benign and 'M' for malignant; all other variables are numeric, and 'id' is not required for making predictions.

1. **Data exploratory**

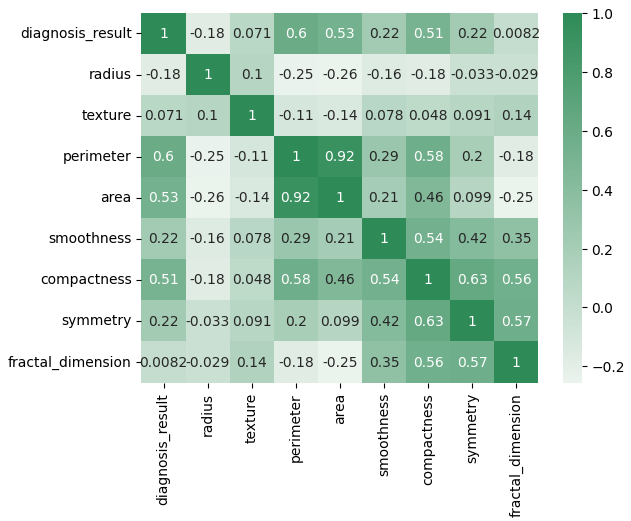
I observed the dataset and found some missing values. Since all the variables are numerical, I decided to choose median as the value to fill in the blank.

From 100 biopsies collected, there are 62 diagnosed as Malignant and 38 as Benign as shown bellow.

Chart, bar chart

Description automatically generated

According to the heatmap below, we can see three variables with the most correlated relationship with diagnosis results are “perimeter”, “area” and “compactness” at 0.6, 0.53, and 0.51 respectively.

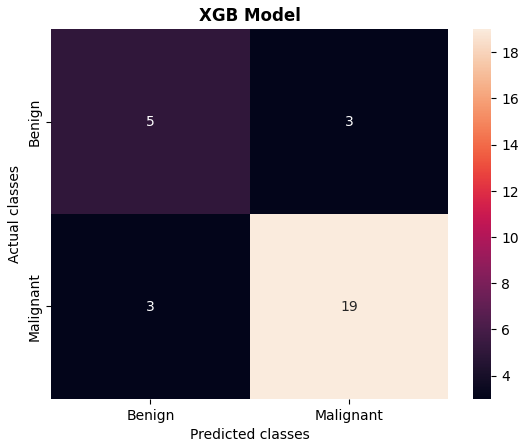


1. **Defining and Fitting Classifier**

The dataset is split into the training dataset and a test dataset with the proportion 70/30

In this project, I used three classifiers: Decision Tree, Random Forest, and XGBoost.

1. **Results**



With 80% accuracy prediction, XGB Model gives the best results. With 3 FN (Malignant but predicted Benign) and 3 FP (Benign but predicted Maligned).

1. **Conclusion**

The project was for practicing building classification models on a dataset. All the attributes are numerical and there was not much effort put into cleaning the data. Since the dataset is relatively small, the result is expected to be not so significant. However, XGB Model performed well with 80% prediction accuracy.

For the purposes of the prediction is to classify Malignant and Benign