SC1015 Mini Project:

Cancer Data

Aanya, Adam, Maeko
FCSH Group 5

TABLE OF CONTENTS

About our DataSet

Utilizing the Cancer Dataset sourced from Kaggle

- The dataset contains mean values of various visual attributes associated with the tumors
- Such as radius, texture, perimeter, area, smoothness, compactness, concavity, and concave points of the tumour
- Unique ID for each patient and classifies tumors as either *Benign (B)* or *Malignant (M)*.

Problem Statement

Are we able to predict accurately whether a tumour is being classified:

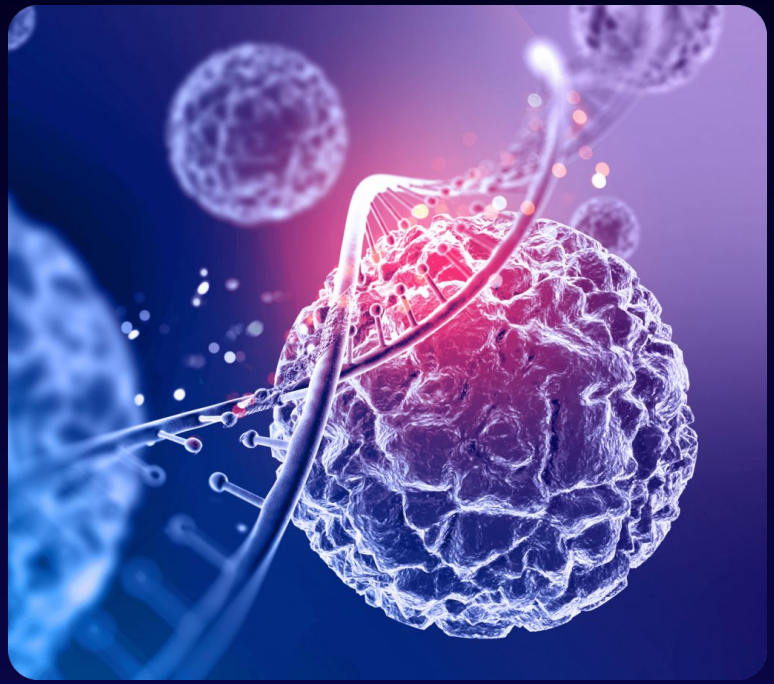
- *Benign (Good Tumor)*
- *Malignant (Bad Tumor)*

based on the variables chosen.

01
Our Motivation

10 MILLION !!!

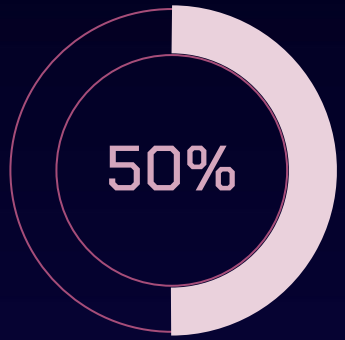
CANCER

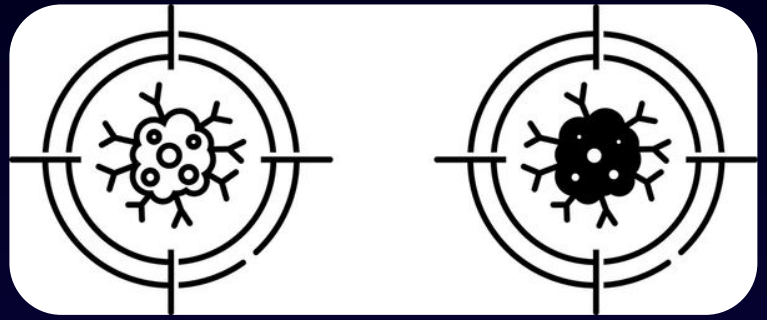


THE IMPORTANCE OF EARLY DETECTION

Diagnosed at the last stage

~50% of cancers are at an advanced stage when diagnosed.





Identifying visual characteristics would allow healthcare providers to develop screening protocols to detect cancer at earlier stages.

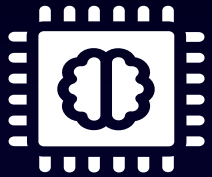
MORE EFFECTIVE TREATMENT

Severity

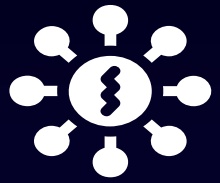
The survival rate of cancer is more than three times higher when the disease is diagnosed early.

02
Setting the stage

Cleaning the data



BENIGN
(Good)



MALIGNANT
(Bad)

Selecting a category to extract variables from	Subcategory division	Choosing the 3 variables
• Mean • SE • ***Worst***	• Between area, perimeter and radius. • We choose ***area*** [similar definition]	Explained on the next slide :)

The 3 Variables

BENIGN
(Good)

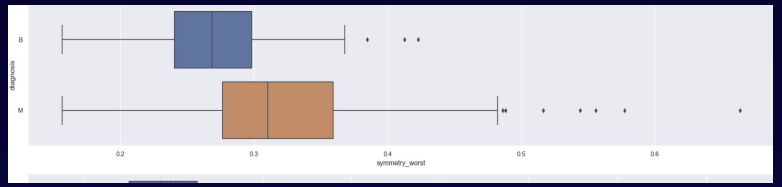
VS

MALIGNANT
(Bad)

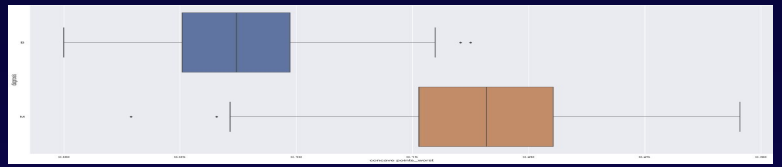
Concavity_points

Concavity_worst

Area_worst



symmetry_worst



concave_points_worst

03

Core Analysis

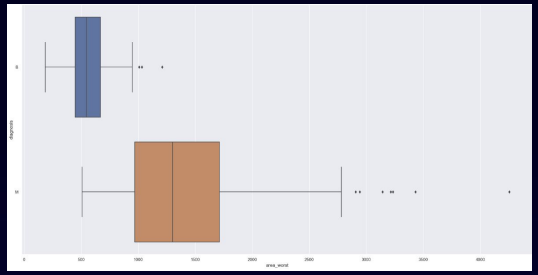
Data Visualisation of the 3 variables

Box Plot

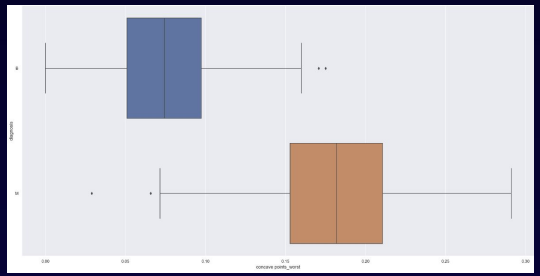
We used a boxplot to clearly visualise the variables namely:

- The difference in parameter
- The greater the difference
  - The stronger the variables impact on predicting M or B

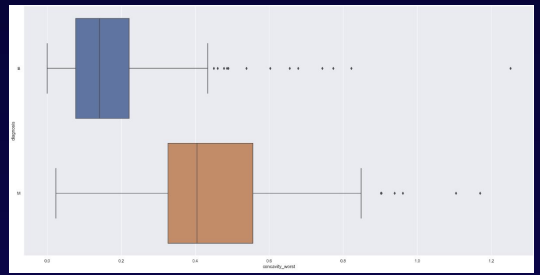
CONCAVITY



CONCAVE POINTS



AREA

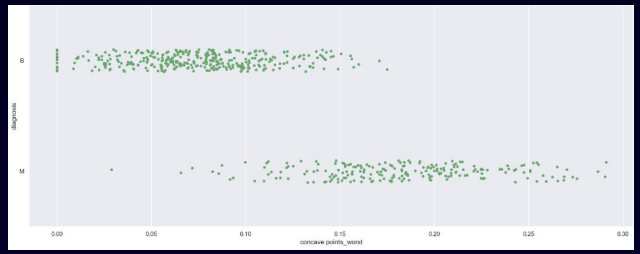


Data Visualisation of the 3 variables

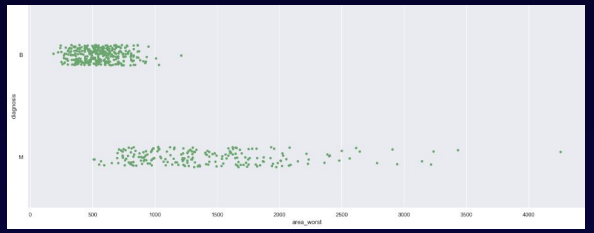
Strip Plot
Helped us visualise the spread of data.

Identify any large sets of anomalies

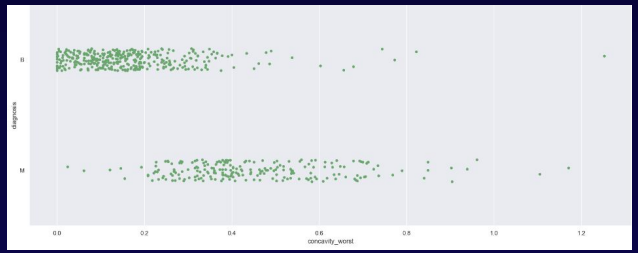
AREA



CONCAVE POINTS



CONCAVITY



04
Machine Learning Model

What have we done?

Uni-Variate Decision Tree

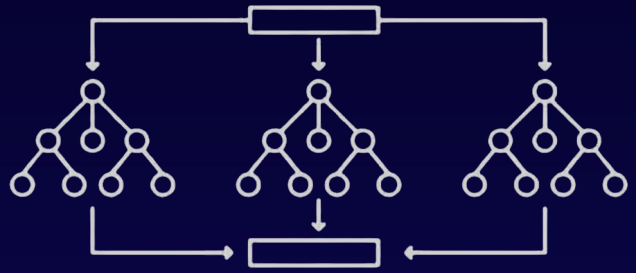
Multi-Variate Decision Tree

Random Forest Classifier (with Cross-Validation)

Our Goal:

- Higher Classification Accuracy
- Lower False Negative Rate (FNR)
- Higher TPR & TNR

Increasing Priority



Our Approach:

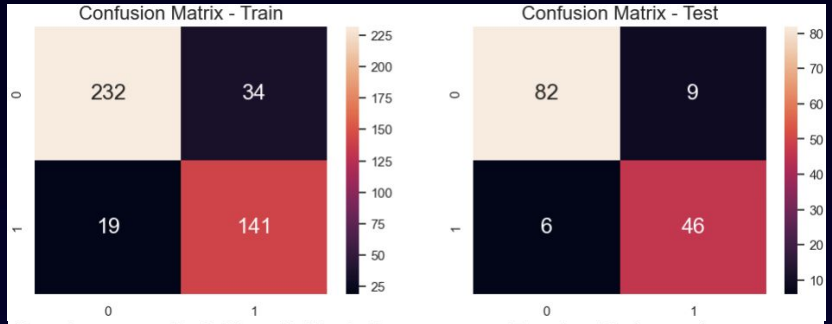
Made use of:

- Decision tree and confusion matrix
- To analyse the relationship of our variables with the diagnosis of either Benign or Malignant

Uni-Variate Decision Tree

Relationship with Concavity

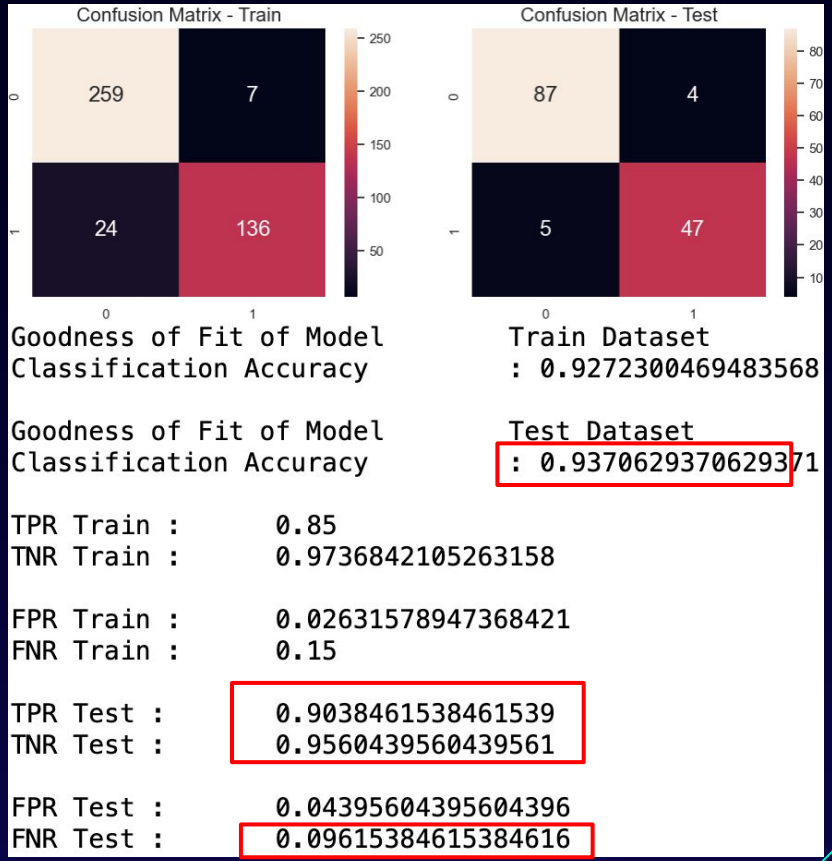
- Worst prediction model
  - *Lowest* Classification Accuracy: 0.895
  - *Lowest* TPR & TNR: 0.885 , 0.901
  - Relatively high FNR: 0.115




```
Goodness of Fit of Model          Train Dataset
Classification Accuracy           : 0.8755868544600939

Goodness of Fit of Model          Test Dataset
Classification Accuracy           : 0.8951048951048951

TPR Train :       0.88125
TNR Train :       0.8721804511278195

FPR Train :       0.12781954887218044
FNR Train :       0.11875

TPR Test :        0.8846153846153846
TNR Test :        0.9010989010989011

FPR Test :        0.0989010989010989
FNR Test :        0.11538461538461539
```

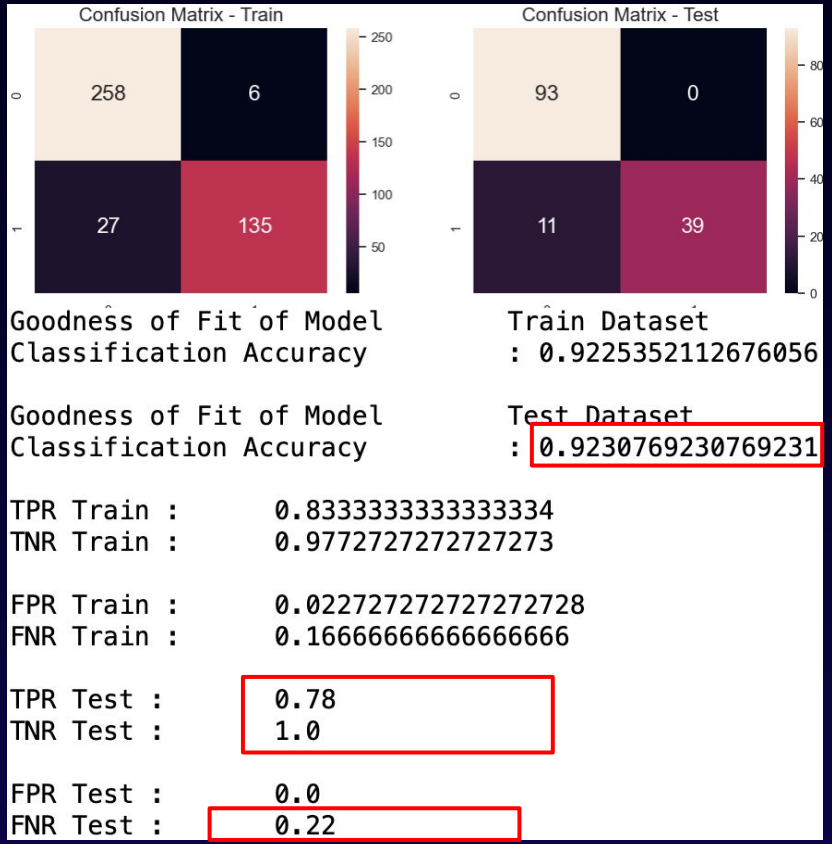
Relationship with Concave_ points

- Highly accurate prediction model
  - *Highest* Classification Accuracy: 0.937
  - *Highest* TPR & TNR: 0.904 , 0.956
  - *Lowest* FNR: 0.096



Relationship with Area

- Fairly good prediction model
  - Classification Accuracy: 0.923
  - TPR & TNR: 0.780 , 1.000
  - *Highest* FNR: 0.220



Comparing all 3 variables (from Test Set)

	Concavity	Concave_Points	Area
Accuracy (highest)	0.895	0.937	0.923
TPR (highest)	0.884	0.903	0.78
TNR (highest)	0.901	0.956	1.0
FNR (lowest)	0.115	0.096	0.22

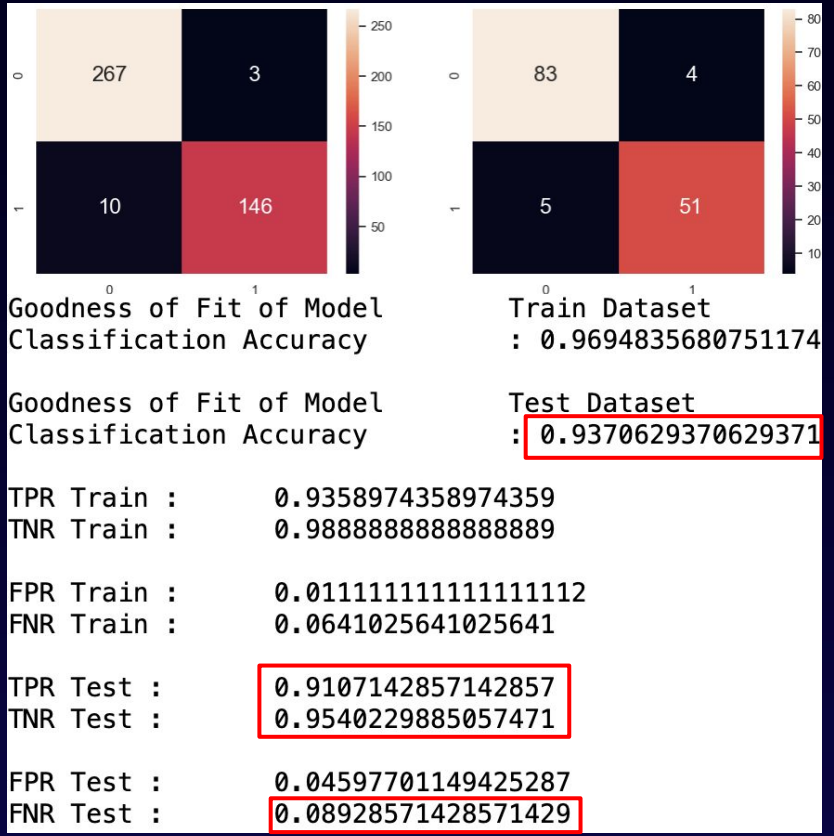
What if we include all 3 variables in a Decision Tree?

Multi-Variate Decision Tree

MultiVariate Comparison

- Better model compared to Uni-variate
  - *Higher* Classification Accuracy: 0.937
  - *Similar* TPR and TNR: 0.911 , 0.954
  - *Lower* FNR: 0.089

Can we do it better?

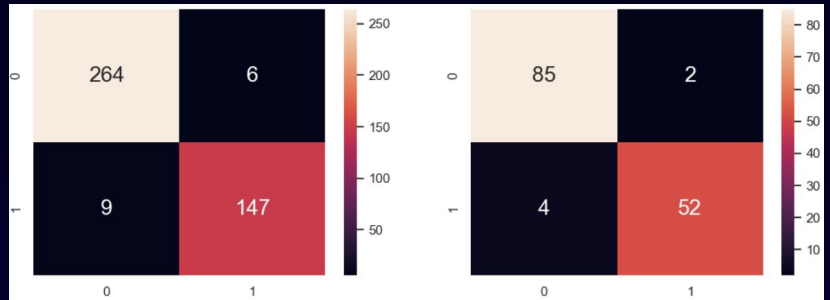


Random Forest Classifier

(with Cross-Validation)

Random Forest Classifier (with Cross-Validation)

- Gives the *best* classification accuracy: 0.958 (the best one yet)
- A *better* TPR and TNR: 0.929 , 0.977
- A much *better* FNR: 0.071




```
Goodness of Fit of Model          Train Dataset
Classification Accuracy           : 0.9647887323943662

Goodness of Fit of Model          Test Dataset
Classification Accuracy           : 0.958041958041958

TPR Train :       0.9423076923076923
TNR Train :       0.9777777777777777

FPR Train :       0.022222222222222223
FNR Train :       0.057692307692307696

TPR Test :        0.9285714285714286
TNR Test :        0.9770114942528736

FPR Test :        0.022988505747126436
FNR Test :        0.07142857142857142
```

04 Conclusion

How our data analysis addressed the problem statement?

- Concavity, area and concave points are good variables to predict whether a tumour is Benign or Malignant.
- This would give patients and doctors the right steps to take if a tumour has been predicted to be bad, which might even save lives.

References

- Erdemtaha. (n.d.). Cancer Data. *Kaggle.* https://www.kaggle.com/datasets/erdemtaha/cancer-data

- The Guardian. (2015, August 10). Cancer survival rates higher with early diagnosis. *The Guardian.* https://www.theguardian.com/society/2015/aug/10/cancer-survival-rates-higher-early-diagnosis

THANK YOU!

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**