Contents

[Section 1: Introduction 2](#_Toc184554859)

[Purpose: 2](#_Toc184554860)

[Goals: 2](#_Toc184554861)

[Section 2: Methodology 2](#_Toc184554862)

[Section 3: Result & Analysis 6](#_Toc184554863)

[Summary of Results: 6](#_Toc184554864)

[Section 4: Conclusions 7](#_Toc184554865)

[Findings: 7](#_Toc184554866)

[Feature Importances: 8](#_Toc184554867)

[Figure 1 Cancer Dataset 2](#_Toc184561503)

[Figure 2 Histograms 4](#_Toc184561504)

[Figure 3 Box Plot 4](#_Toc184561505)

[Figure 4 Statistical Summary 5](#_Toc184561506)

[Figure 5 Features & Label 5](#_Toc184561507)

[Figure 6 Standardize Features (x=data point, µ=data series mean, sigma=std dev) 5](#_Toc184561508)

[Figure 7 Split Dataset into Features & Target followed by Training & Testing Data 5](#_Toc184561509)

[Figure 8 Verify Standardizing of data 5](#_Toc184561510)

[Figure 9 Predicted Results Random Forest Regression 7](#_Toc184561511)

[Figure 10 Predicted Results Random Forest Classifier 7](#_Toc184561512)

[Figure 11 Features Importances in Ascending Order 8](#_Toc184561513)

[Figure 12 Decision Boundary for texture\_mean & radius\_mean 9](#_Toc184561514)

[Figure 13 Decision Boundary for perimeter\_worst & area\_worst 9](#_Toc184561515)

[Figure 14 Correlation Matrix Heat Map for all Features 10](#_Toc184561516)

[Figure 15 ROC CURVE 10](#_Toc184561517)

[Table 1 Accuracies for 3 different parameters for Random Forest Regression 6](#_Toc183973470)

[Table 2 Accuracies for 3 different parameters for Random Forest Classification 6](#_Toc183973471)

# Section 1: Introduction

Purpose: Create & compare different ML (Machine Learning) algorithms for generating models for predicting Cancer Data. This dataset contains the characteristics of patients diagnosed with cancer.

Figure 1 Cancer Dataset

A screenshot of a computer

Description automatically generated

## Goals:

This dataset will be used to train & test models and algorithms used to make cancer diagnoses. The Label for Dataset is highlighted in Red as shown in Figure *1* Cancer Dataset. The Label indicates if the disease is M (Malignant for positive diagnosis) or B (Benign for negative diagnosis). Early diagnosis is key to surviving cancer & ML gives the ability to analyze complex patterns.

# Section 2: Methodology

1. Load datasets from sklearn into Data Frame.
2. Analyze the dataset by plotting some histograms as shown in *Figure 2* Histograms & a boxplot shown in *Figure 3* Box Plot for the features of dataset along with a summary as shown in Figure 4 Statistical Summary.
3. Histograms show us the distribution of numerical data for each feature, while the box plot shows a graphical representation of the mean, median, standard deviation & outliers for each feature in dataset.
4. The statistical summary tabulates the box plot results in an easy-to-read description.
5. Run the dataset through function “removeCategories(df)” to separate out features containing categorical values, which are basically of data type “String” from numerical features, because categorical features data will need to be converted to numerical ones for modelling & training. This function returns a tuple of (numerical features, categorical features).
6. As seen in Figure *1* Cancer Dataset, the target Label “diagnosis” is the only categorical data series. See Figure 5 Features & Label, for all Features & the Label.
7. Extract all the features from Dataset & store in Data frame X. Extract the Label into Data series y.
8. Split X & y from step 7 into Training & Testing data sets, by allocating 24% of data to testing. Graphical representation of this is shown in *Figure 7* Split Dataset into Features & Target followed by Training & Testing Data.
9. Split the data set as shown in Figure 7 Split Dataset into Features & Target followed by Training & Testing Data.
10. Inherently features are characteristics of the entity being modelled & hence different features may have different units of measurement & weights. We will be using regression algorithms which will work better if we scale or standardize the feature data which will result in better accuracy & consistency of the predictive model, since all features will be brought to same format or units.
11. Standardize all features from all data sets as shown in Figure 6 Standardize. Ensure that standardization has been applied correctly by validating that the mean of each standardized feature data series amounts to 0 & the standard deviation is 1, as shown in Figure 8 Verify Standardizing of data.
12. As explained in step 5 above, we will convert categorical target Label “diagnosis” to numerical values using python library “LabelEncoder”, so it can used effectively in regression modelling.
13. Send Training Data set into ML algorithm “RandomForestRegressor” for modelling using parameters of 100 trees in forest & randomness value of 42 for bootstrapping of the samples used when building trees. This will generate our model after fitting data.
14. We will then predict our testing data from Step 8 & 9 above to yield predicted data. This result will be numerical in nature (between 0 & 1 real), since we assigned Label of 1 to Malignant & 0 to Benign.
15. Predicted data in step 14 needs to be Discretized so it can be compared to categorical Target Label of 1(Malignant) or 0(Benign), so we can computer accuracy.
16. Following up from Step 15 above, we will use a function “discretize(model, y\_predict)” & set a threshold value of 0.4(adjustable) which is the cut-off above which the disease is Malignant.
17. As shown in the RMSE = 0.16. We can also compute the accuracy on the discretized prediction & Accuracy score using Random Forest Regression Algorithm for: Cancer Data is: 0.99
18. We will perform a Classifier Algorithm using Random Forest modelling with the same parameters as described in step 13 above & repeat the fitting & predictions of testing data. Accuracy score using Random Forest Classifier Algorithm for: Cancer Data is: 0.98.

Figure Histograms

A screenshot of a graph

Description automatically generated

Figure Box Plot

A graph with text on it

Description automatically generated

Figure Statistical Summary

A screenshot of a computer

Description automatically generated

Figure Features & Label

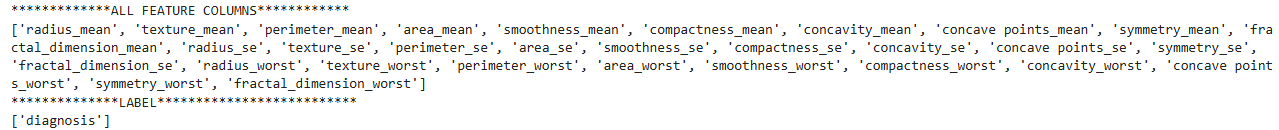


Figure Standardize Features (x=data point, µ=data series mean, sigma=std dev)



Figure Split Dataset into Features & Target followed by Training & Testing Data

A diagram of a train test

Description automatically generated with medium confidence

Figure Verify Standardizing of data

A white background with black text

Description automatically generated

# Section 3: Result & Analysis

### Summary of Results:

Table Accuracies for 3 different parameters for Random Forest Regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Random Forest Regression** | **Accuracy** | **RMSE** | **Parameters** | |
| **Cancer Data Set Tuning** |  |  | **# of trees** | **random\_state** |
|  |  |  |  |  |
| **1** | 0.99 | 0.16 | 100 | 42 |
| **2** | 0.97 | 0.17 | 50 | 42 |
| **3** | 0.98 | 0.16 | 500 | 42 |

Table Accuracies for 3 different parameters for Random Forest Classification

|  |  |  |  |
| --- | --- | --- | --- |
| **Random Forest Classifier** | **Accuracy** | **Parameters** | |
| **Cancer Data Set Tuning** |  | **# of trees** | **random\_state** |
|  |  |  |  |
| **1** | 0.98 | 100 | 42 |
| **2** | 0.97 | 50 | 42 |
| **3** | 0.98 | 500 | 42 |

Parameters chosen:

Table 1 Accuracies for 3 different parameters for Random Forest Regression shows tuning .

Table 2 Accuracies for 3 different parameters for Random Forest Classification shows tuning.

Large values of # of trees does not have an effect on accuracy.

Refer Figure 15 ROC CURVE on page 10, for model to analyze the trade-off between True Positive Rate & False Positive Rate.

# Section 4: Conclusions

## Findings:

1. Either Random Forest Regression or Random Forest Classification algorithm can be used for the cancer dataset as the accuracies are close to each other. Additionally increasing the size of forest beyond 100 does not improve accuracy.
2. Random Forest Classification algorithm gives similar accuracy results as Random Forest Regression for simpler model algorithm. The classification algorithm is binary in nature (1 or 0, M or B) & hence does not have to be discretized for comparison or prediction after modelling. See Figure 9 Predicted Results Random Forest Regression & Figure 10 Predicted Results Random Forest Classifier.
3. However, following up on Step 2 above, the Regression model can yield more control over the accuracy since we can control the threshold value (currently set at 0.4). This can be increased to reduce false positives or reduced to provide a more aggressive or pessimistic diagnosis to ensure early detection.

Figure Predicted Results Random Forest Regression

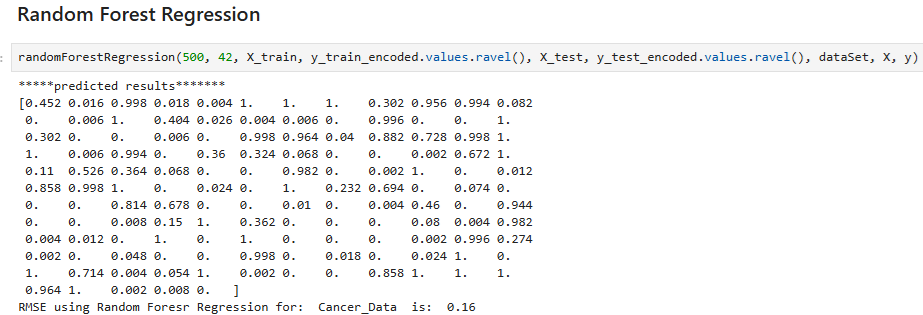




Figure Predicted Results Random Forest Classifier

A screenshot of a computer

Description automatically generated

## Feature Importances:

1. Refer Figure 11 Features Importances in Ascending Order for the features that contribute most to the Target outcome which is basically Cancer diagnosis of Malignant or Benign.
2. Refer Figure 12 Decision Boundary for texture\_mean & radius\_mean for a Decision Boundary between the first and second features “texture\_mean” & “radius\_mean”. Blue region is Benign & Red region is Malignant diagnosis.
3. Area\_worst & perimeter\_worst are the features that top this list. Hence will retrain Random Forest classification with these features & plot the decision boundary.
4. See Figure 13 Decision Boundary for perimeter\_worst & area\_worst. Blue region is Benign & Red region is Malignant diagnosis. We can observe that for higher values of perimeter\_worst & area\_worst (greater than 0 for both the diagnosis is Malignant)
5. Refer Figure 14 Correlation Matrix Heat Map for all Features, which shows relationship between features in data set. Area\_worst & perimeter\_worst have a correlation coefficient of 1, implying a highly positive linear relationship. Hence the points on decision boundary as per Figure 13 Decision Boundary for perimeter\_worst & area\_worst are in a straight line.
6. Refer Figure 14 Correlation Matrix Heat Map for all Features. Features radius\_mean & texture\_mean have a correlation coefficient of 0.3, implying no linear relationship exists. Hence the points on decision boundary as per Figure 12 Decision Boundary for texture\_mean & radius\_mean, are scattered all over the plot.

Figure Features Importances in Ascending Order

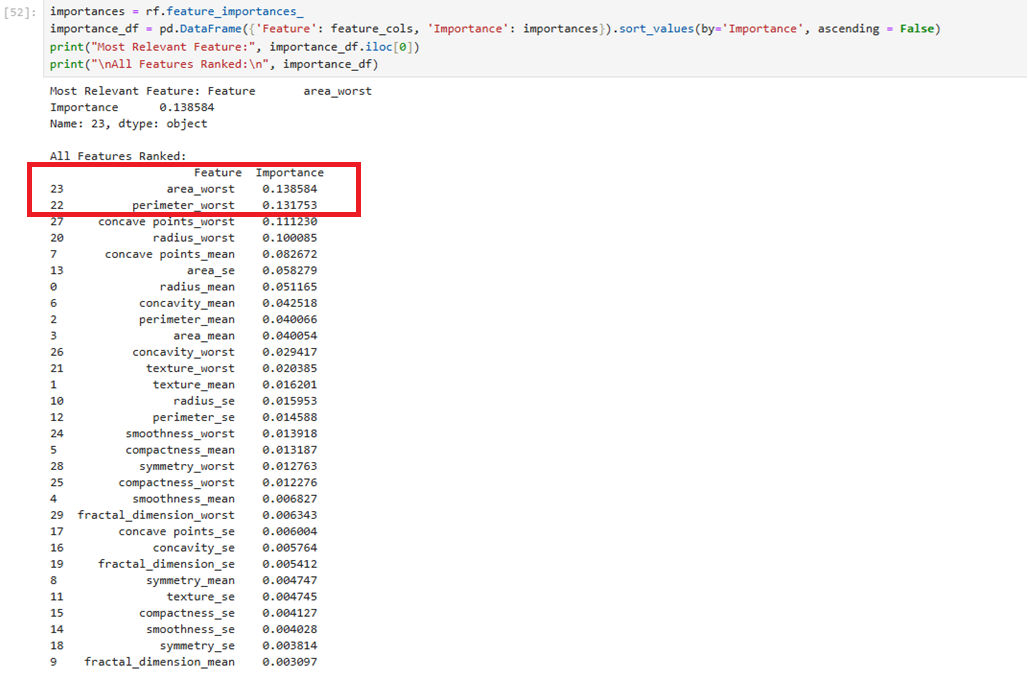


Figure Decision Boundary for texture\_mean & radius\_mean

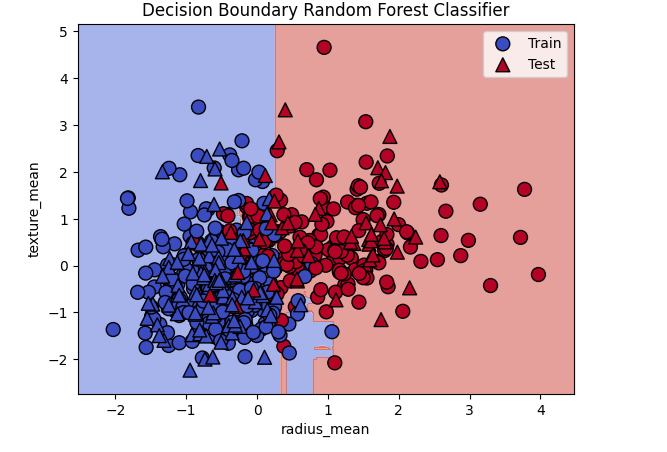


Figure Decision Boundary for perimeter\_worst & area\_worst

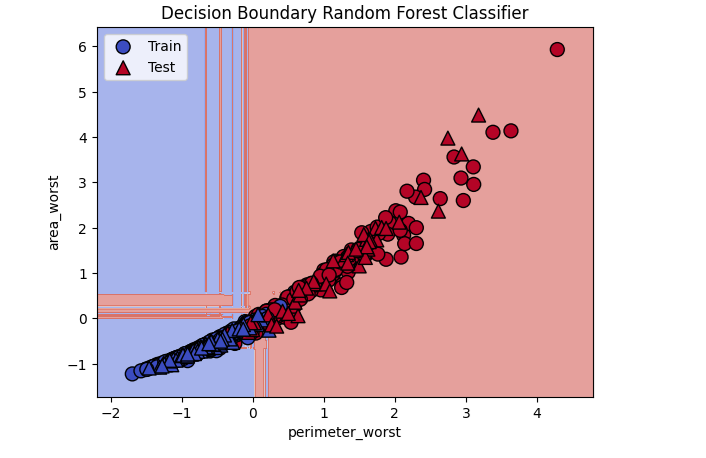


Figure Correlation Matrix Heat Map for all Features

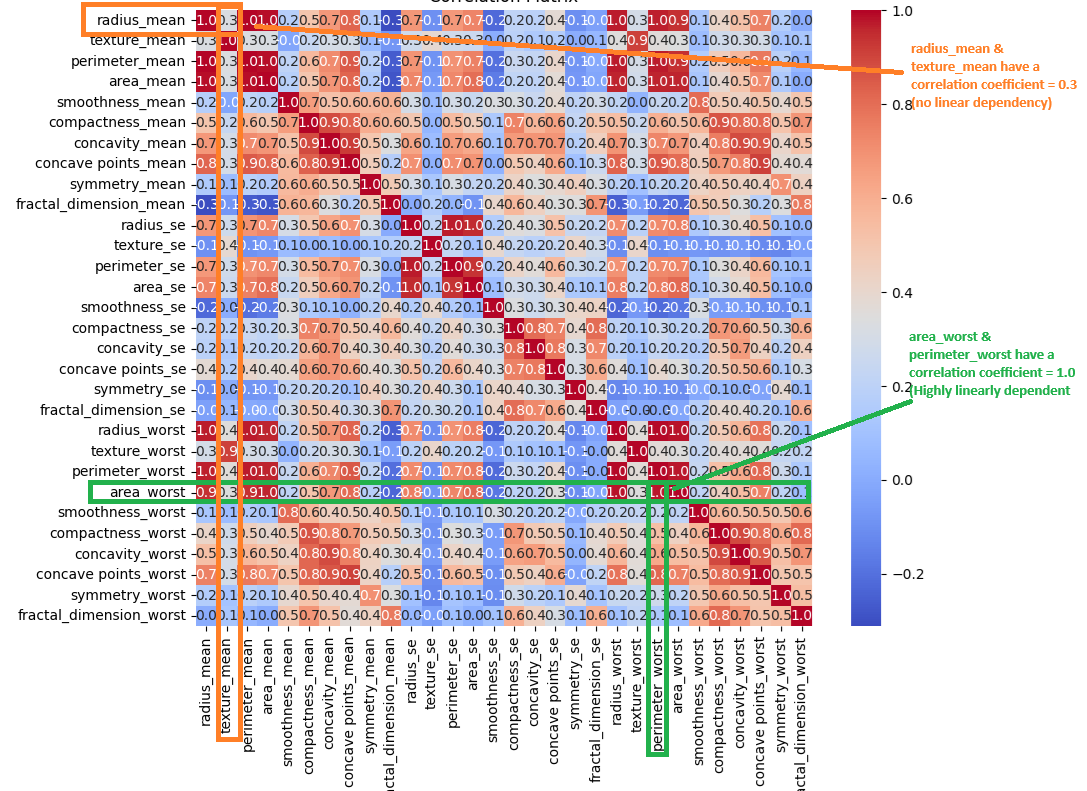


Figure ROC CURVE

A graph with a red line

Description automatically generated