Stat-581: Probability and Statistical Inference for Data Science
Project Report

Keya Desai (kd706)
Twisha Naik (tn268)
Prakruti Joshi (phj15)

Abstract

Breast Cancer Detection: Given the data of a tumor (breast mass), the task is to classify whether the mass is malignant (cancerous) or benign (non-cancerous).

Dataset - Breast Cancer Wisconsin (Diagnostic) Dataset

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Attribute Information:

1) ID number
2) Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

1) radius (mean of distances from center to points on the perimeter)
2) texture (standard deviation of gray-scale values)
3) perimeter
4) area
5) smoothness (local variation in radius lengths)
6) compactness (perimeter^2 / area - 1.0)
7) concavity (severity of concave portions of the contour)
8) concave points (number of concave portions of the contour)
9) symmetry
10) fractal dimension ("coastline approximation" - 1)

The mean, standard error and “worst” or largest (mean of the three largest values) of these features were computed for each image, **resulting in 30 features.** For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

No. of data points = 569
No. of features = 31
Class distribution: 357 benign, 212 malignant

Understanding the data

1. *Radius* - Mean of distances from center to points on the perimeter
2. *Texture* - Standard deviation of gray-scale values
3. *Perimeter*
4. *Area*
5. *Smoothness* - How smooth are the edges of the mass. It is the local variation in radius lengths. If all the points on the perimeter are equally far from the center, it will be smooth and hence the value of this parameter will be less.
6. $Compactness = \frac{perimeter^2}{area} - 1.0$

This defines how compact is the given mass. A perfect circle would have the maximum compactness as it will have the maximum area for a given perimeter.

7. *Concavity* - As shown in Figure 1, there might be multiple concave points in the datasets. This feature describes the severity of concave portions of the contour.

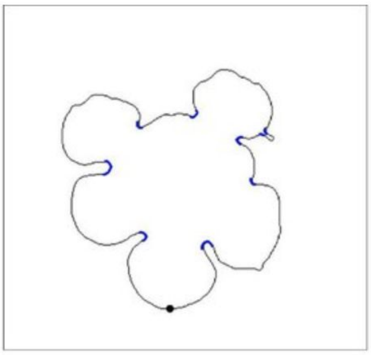


Figure 1. Sample Image of the tumour

8. *Concave points* - Number of concave portions of the contour
9. *Symmetry*
10. *Fractal dimension* - The precision with which the parameters are measured. For example, we have a non-smooth boundary for the given mass and we want to find the perimeter. If we increase the precision of measurement used, we get a difference in the measured perimeter.

Data Analysis

Since the mean, standard error and the mean of the 10 features have been taken, we suspect that they might be highly correlated. Hence, we start the data analysis with looking at the correlation matrix.

1. Correlation

The correlation between two features x and y can be defined by the following formula:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\surd \sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2} \qquad (1)$$

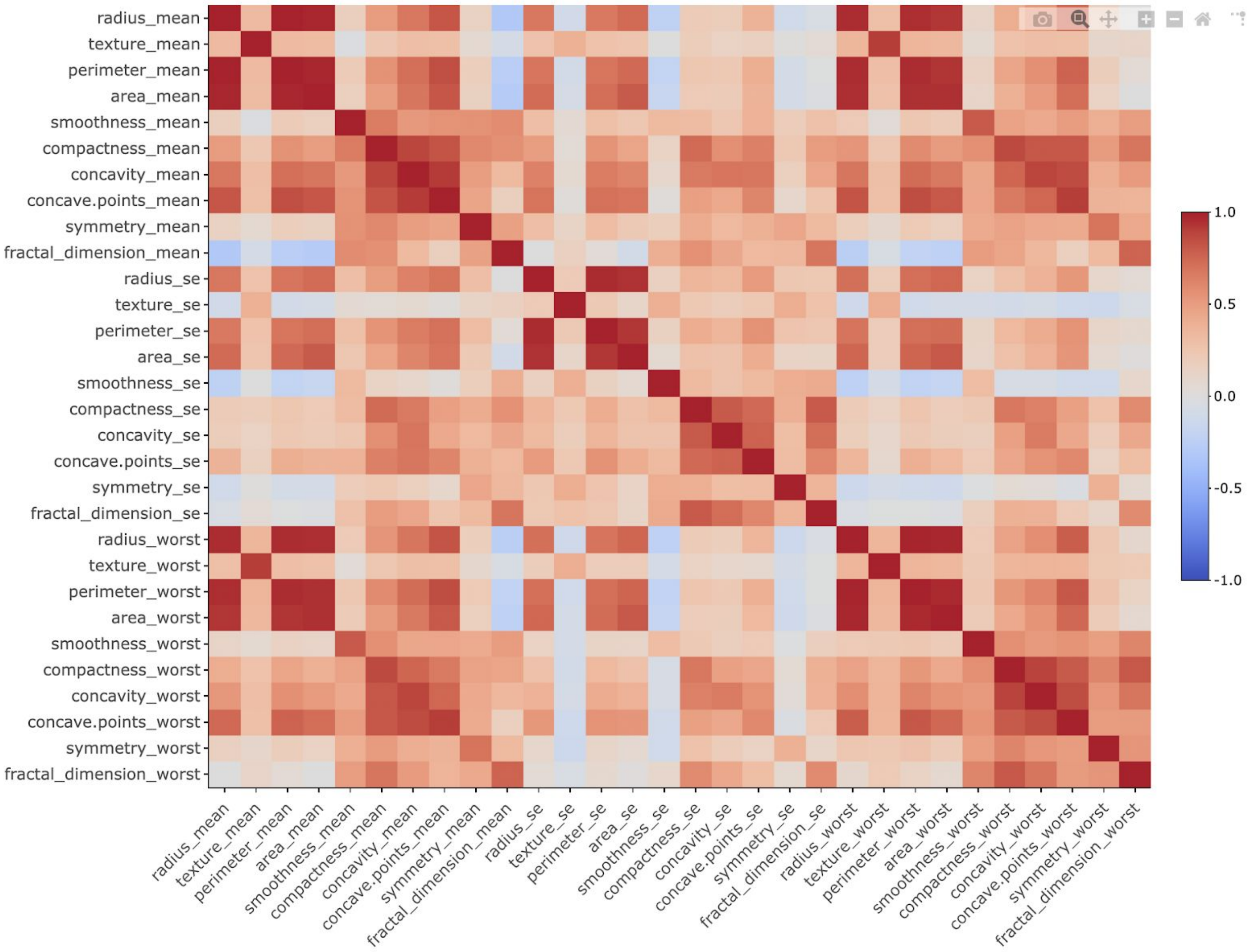


Figure 2. Correlation matrix of the 30 features in the data

From Figure 2., we can observe that there is a high correlation between the mean, standard error and the worst of each feature. For instance, radius mean, radius standard error and radius worst are highly correlated. To understand the relation between the 10 underlying features, we plot the correlation matrix of means of each of these.

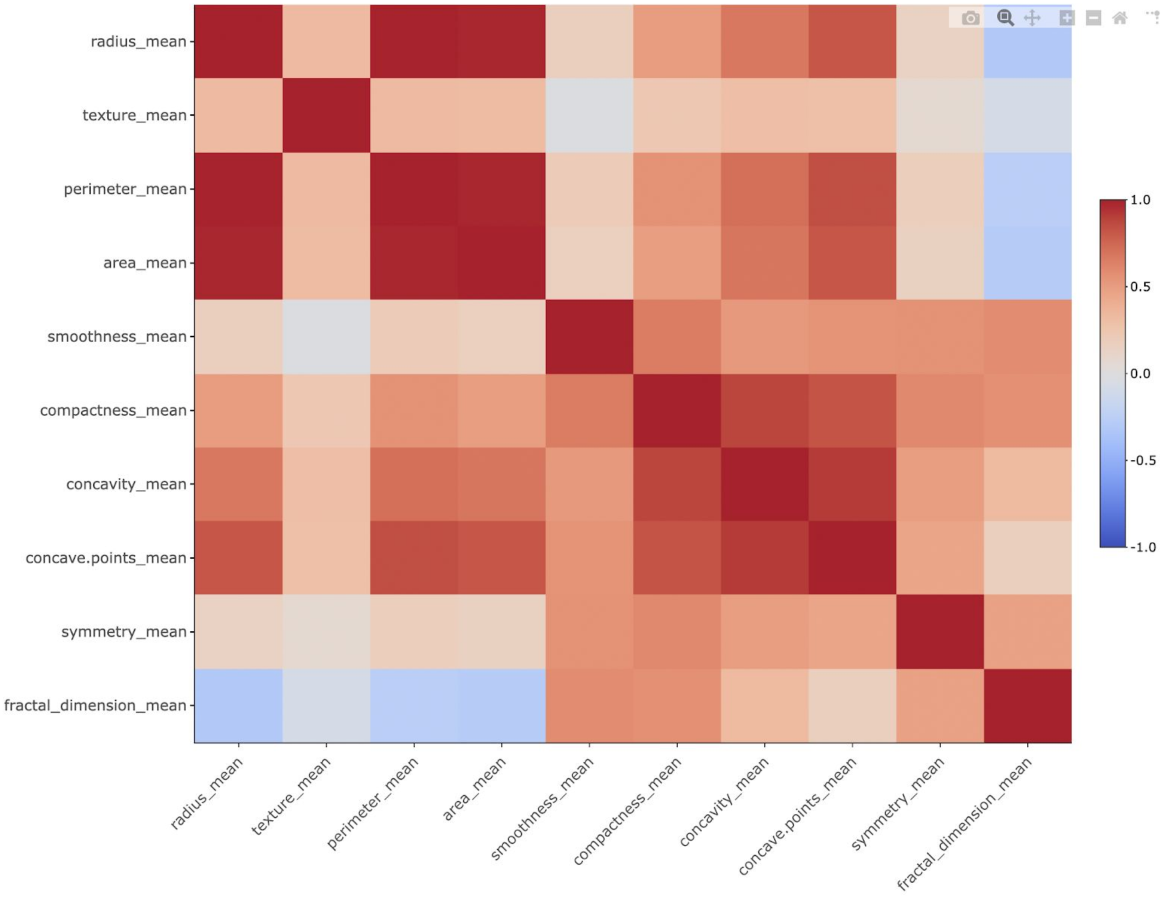


Figure 3. Correlation of the means of the underlying features

From Figure 3., we observe that

- Perimeter and Area are highly correlated with radius.
- Compactness, concavity and number of concave points are positively correlated.

Next, we look at the distribution of each of these features categorized by cancer type - Benign or Malignant.

2. Distribution of features

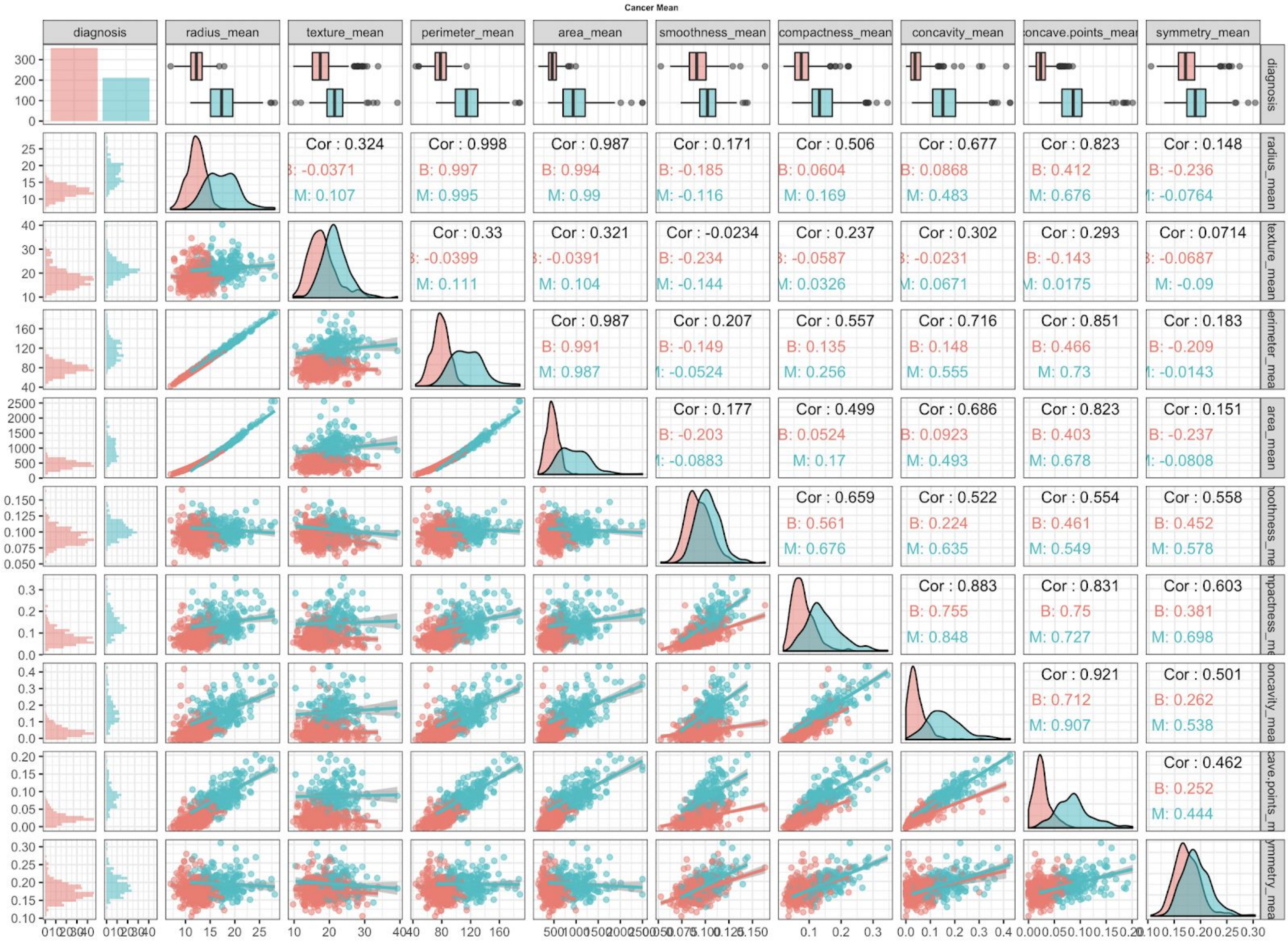


Figure 4. Distribution of the means of the underlying features by cancer type

From the distribution of the features, it seems that the tumor type can be distinguished based on the values of the features itself. This inspires us to try decision trees and random forests as classifiers.

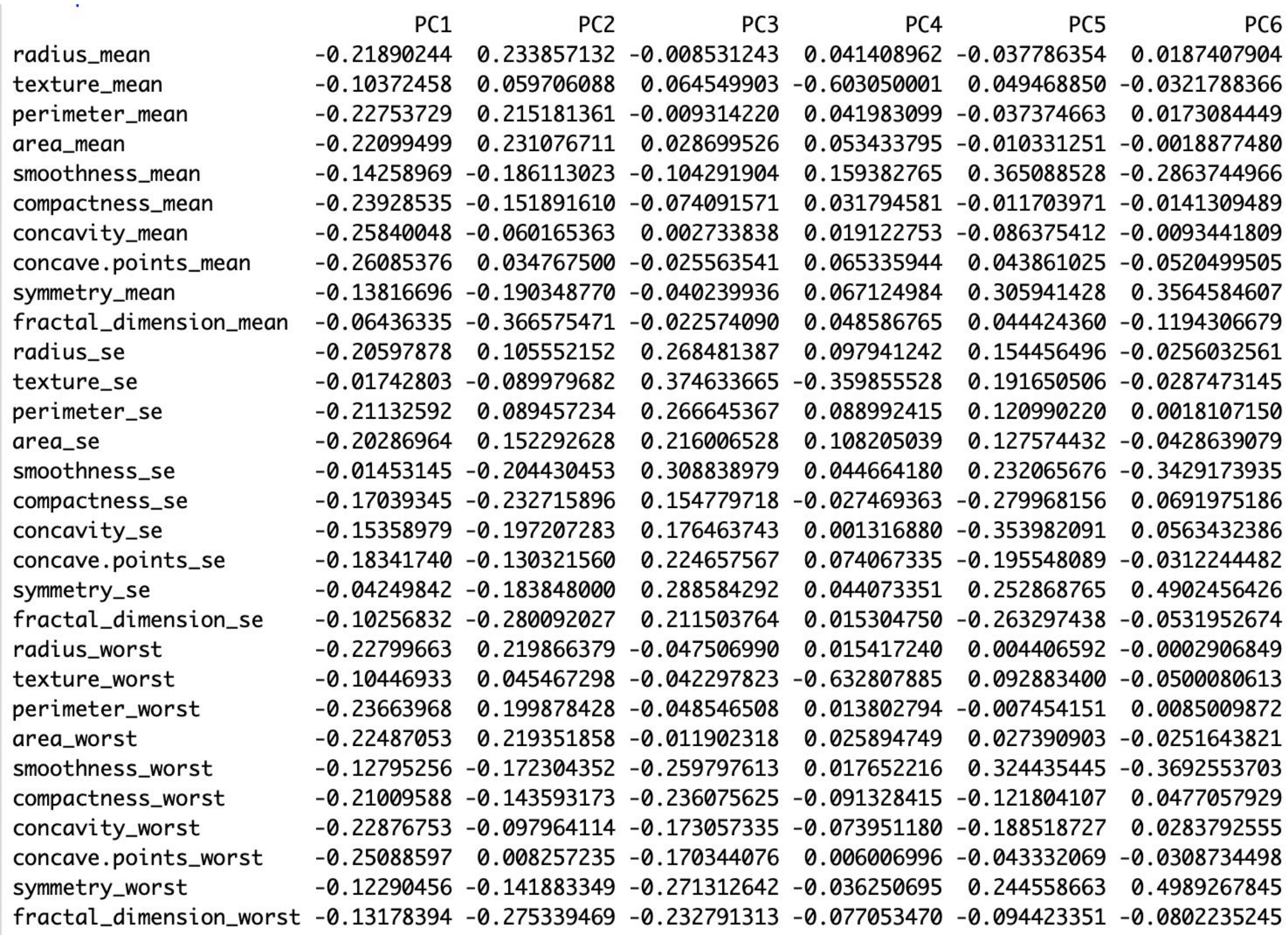
As there are highly correlated features, we next do Principal Component Analysis to reduce the number of features.

3. Principal Component Analysis

Principal Component Analysis (PCA) is a technique to reduce the number of features of the data while still capturing the essence of entire data. After performing PCA, we can describe our data in the form of Principal Components which are linear combination of the original features.

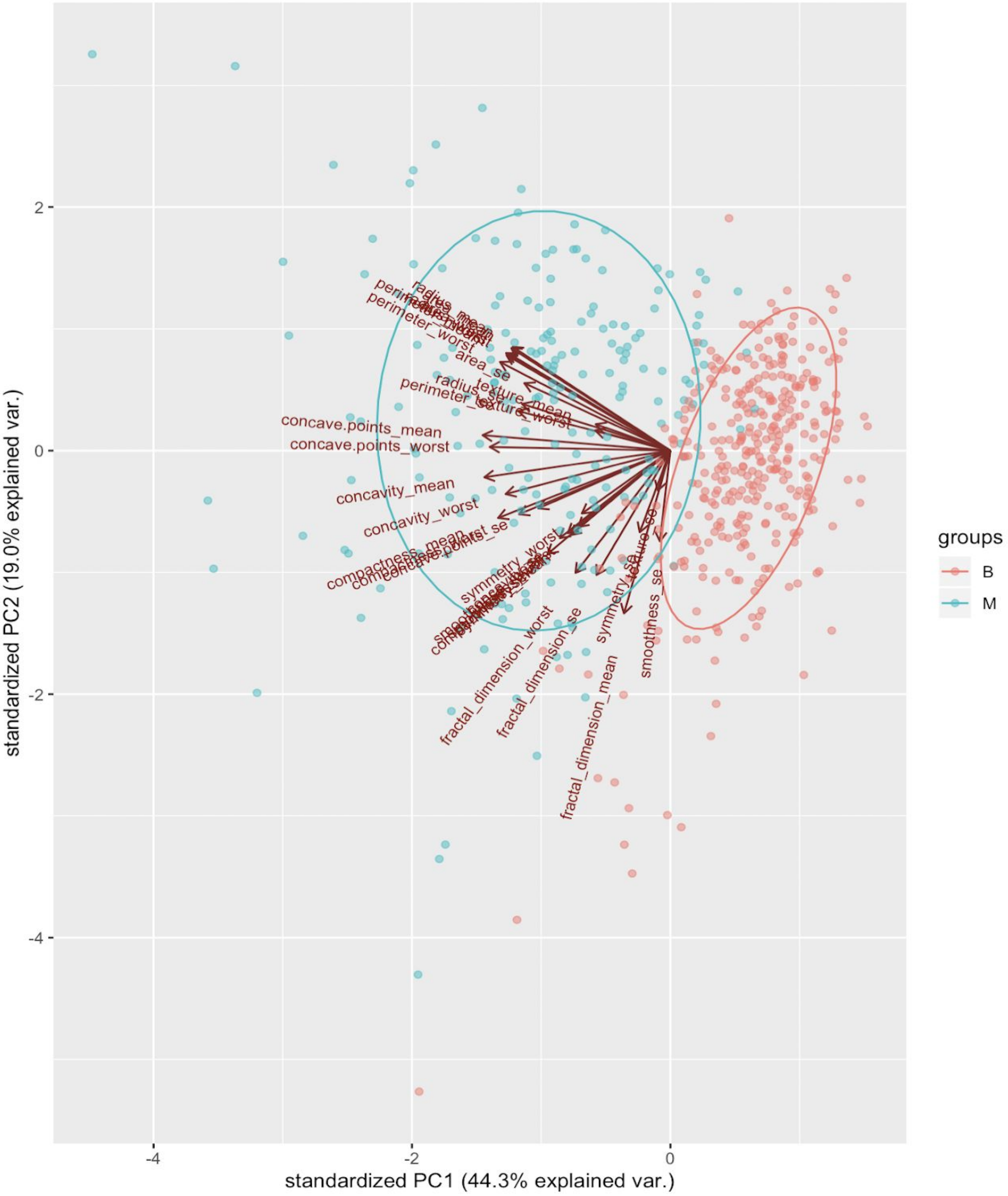
Variation in the feature describes the amount of information contained in it. The goal of PCA is to identify directions (or principal components) along which the variation in the data is maximal.

Interpretation of PCA output:
From the image below, we can see that PC1 is formed by the linear combination of all the features with the specified coefficients. Same follows for the rest of the principal components.

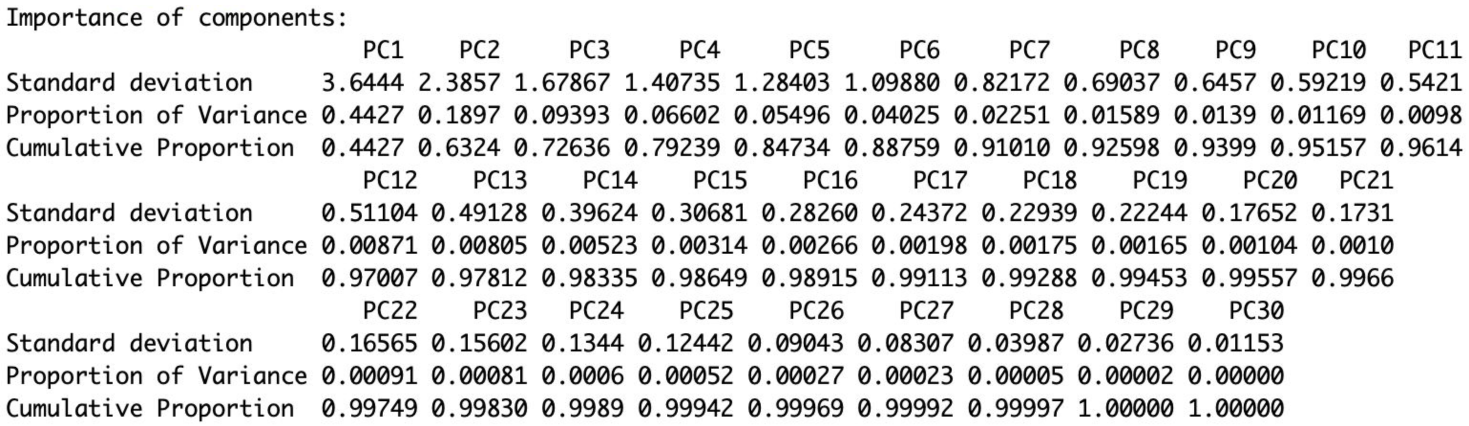


```
                               PC1          PC2          PC3          PC4          PC5           PC6
radius_mean            -0.21890244  0.233857132 -0.008531243  0.041408962 -0.037786354  0.0187407904
texture_mean           -0.10372458  0.059706088  0.064549903 -0.603050001  0.049468850 -0.0321788366
perimeter_mean         -0.22753729  0.215181361 -0.009314220  0.041983099 -0.037374663  0.0173084449
area_mean              -0.22099499  0.231076711  0.028699526  0.053433795 -0.010331251 -0.0018877480
smoothness_mean        -0.14258969 -0.186113023 -0.104291904  0.159382765  0.365088528 -0.2863744966
compactness_mean       -0.23928535 -0.151891610 -0.074091571  0.031794581 -0.011703971 -0.0141309489
concavity_mean         -0.25840048 -0.060165363  0.002733838  0.019122753 -0.086375412 -0.0093441809
concave.points_mean    -0.26085376  0.034767500 -0.025563541  0.065335944  0.043861025 -0.0520499505
symmetry_mean          -0.13816696 -0.190348770 -0.040239936  0.067124984  0.305941428  0.3564584607
fractal_dimension_mean -0.06436335 -0.366575471 -0.022574090  0.048586765  0.044424360 -0.1194306679
radius_se              -0.20597878  0.105552152  0.268481387  0.097941242  0.154456496 -0.0256032561
texture_se             -0.01742803 -0.089979682  0.374633665 -0.359855528  0.191650506 -0.0287473145
perimeter_se           -0.21132592  0.089457234  0.266645367  0.088992415  0.120990220  0.0018107150
area_se                -0.20286964  0.152292628  0.216006528  0.108205039  0.127574432 -0.0428639079
smoothness_se          -0.01453145 -0.204430453  0.308838979  0.044664180  0.232065676 -0.3429173935
compactness_se         -0.17039345 -0.232715896  0.154779718 -0.027469363 -0.279968156  0.0691975186
concavity_se           -0.15358979 -0.197207283  0.176463743  0.001316880 -0.353982091  0.0563432386
concave.points_se      -0.18341740 -0.130321560  0.224657567  0.074067335 -0.195548089 -0.0312244482
symmetry_se            -0.04249842 -0.183848000  0.288584292  0.044073351  0.252868765  0.4902456426
fractal_dimension_se   -0.10256832 -0.280092027  0.211503764  0.015304750 -0.263297438 -0.0531952674
radius_worst           -0.22799663  0.219866379 -0.047506990  0.015417240  0.004406592 -0.0002906849
texture_worst          -0.10446933  0.045467298 -0.042297823 -0.632807885  0.092883400 -0.0500080613
perimeter_worst        -0.23663968  0.199878428 -0.048546508  0.013802794 -0.007454151  0.0085009872
area_worst             -0.22487053  0.219351858 -0.011902318  0.025894749  0.027390903 -0.0251643821
smoothness_worst       -0.12795256 -0.172304352 -0.259797613  0.017652216  0.324435445 -0.3692553703
compactness_worst      -0.21009588 -0.143593173 -0.236075625 -0.091328415 -0.121804107  0.0477057929
concavity_worst        -0.22876753 -0.097964114 -0.173057335 -0.073951180 -0.188518727  0.0283792555
concave.points_worst   -0.25088597  0.008257235 -0.170344076  0.006006996 -0.043332069 -0.0308734498
symmetry_worst         -0.12290456 -0.141883349 -0.271312642 -0.036250695  0.244558663  0.4989267845
fractal_dimension_worst -0.13178394 -0.275339469 -0.232791313 -0.077053470 -0.094423351 -0.0802235245
```

PCA on all features:



Here, as described in the plot, explained variance of PC1 is 44.3% and that of PC2 is 19.0%. Also, we can see the features whose linear combinations result in those principal components.



```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8    PC9    PC10   PC11
Standard deviation     3.6444 2.3857 1.67867 1.40735 1.28403 1.09880 0.82172 0.69037 0.6457 0.59219 0.5421
Proportion of Variance 0.4427 0.1897 0.09393 0.06602 0.05496 0.04025 0.02251 0.01589 0.0139 0.01169 0.0098
Cumulative Proportion  0.4427 0.6324 0.72636 0.79239 0.84734 0.88759 0.91010 0.92598 0.9399 0.95157 0.9614
                          PC12    PC13    PC14    PC15    PC16    PC17    PC18    PC19    PC20   PC21
Standard deviation     0.51104 0.49128 0.39624 0.30681 0.28260 0.24372 0.22939 0.22244 0.17652 0.1731
Proportion of Variance 0.00871 0.00805 0.00523 0.00314 0.00266 0.00198 0.00175 0.00165 0.00104 0.0010
Cumulative Proportion  0.97007 0.97812 0.98335 0.98649 0.98915 0.99113 0.99288 0.99453 0.99557 0.9966
                          PC22    PC23   PC24    PC25    PC26    PC27    PC28    PC29    PC30
Standard deviation     0.16565 0.15602 0.1344 0.12442 0.09043 0.08307 0.03987 0.02736 0.01153
Proportion of Variance 0.00091 0.00081 0.0006 0.00052 0.00027 0.00023 0.00005 0.00002 0.00000
Cumulative Proportion  0.99749 0.99830 0.9989 0.99942 0.99969 0.99992 0.99997 1.00000 1.00000
```

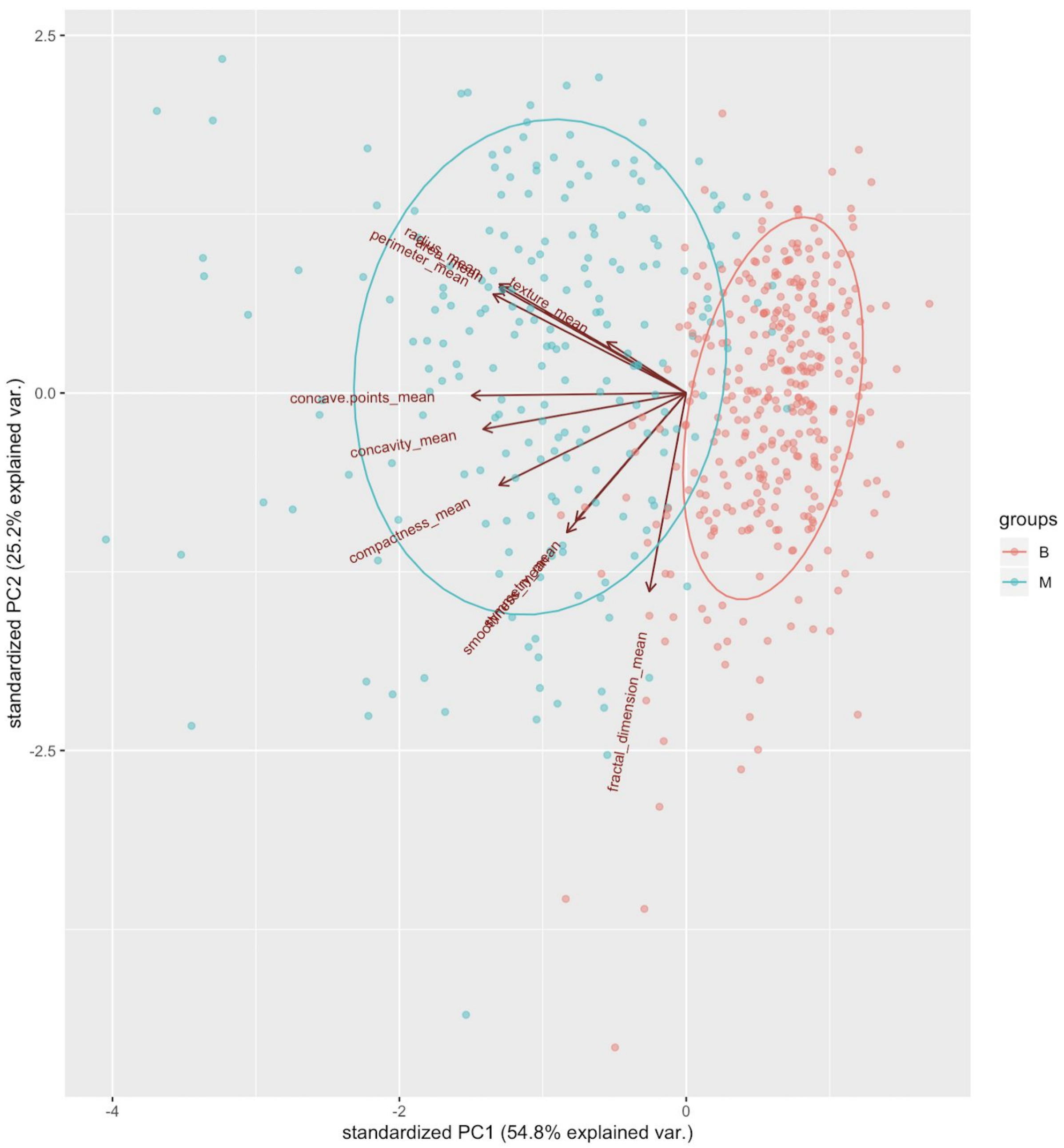
Observations from the PCA output:

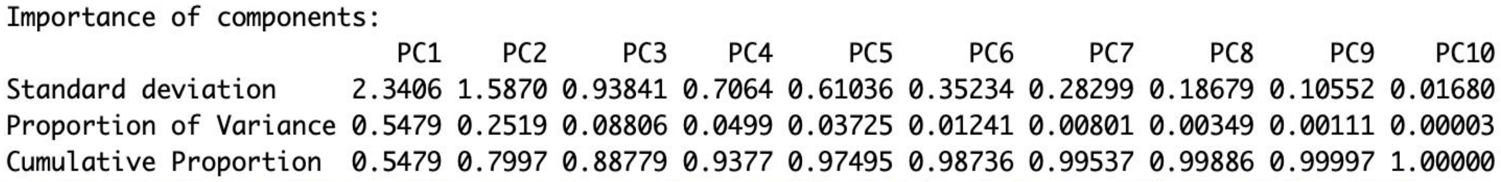
- Number of PCs <= Number of original features (30)
- The cumulative proportion describes the percentage of information captured by the components till the given PC.
- PC1 and PC2 cover the majority of the information in the original data.
- If all the 30 PCs are included, 100% of the information is captured and there is no loss of data.

We also applied PCA to the subset of the 30 features so as to simplify the visualisation and interpretation of the results.

The results for PCA on just the 10 features corresponding to the mean are added below. As seen from the figure, highly correlated features like radius_mean, area_mean and perimeter_mean almost overlap each other when visualized in the PC1 and PC2 plane.

PCA on mean of the features (the first 10 features):





```
Importance of components:
                          PC1    PC2     PC3    PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     2.3406 1.5870 0.93841 0.7064 0.61036 0.35234 0.28299 0.18679 0.10552 0.01680
Proportion of Variance 0.5479 0.2519 0.08806 0.0499 0.03725 0.01241 0.00801 0.00349 0.00111 0.00003
Cumulative Proportion  0.5479 0.7997 0.88779 0.9377 0.97495 0.98736 0.99537 0.99886 0.99997 1.00000
```

Classification

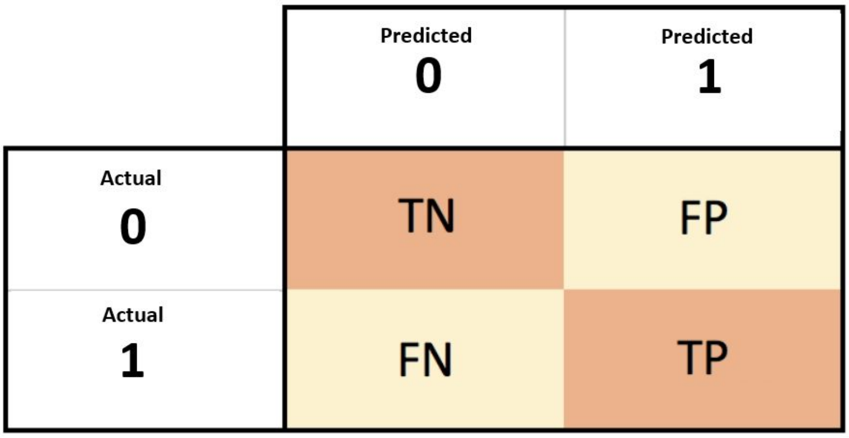
Classification methods tried:

1. Logistic Regression
2. Naive Bayes
3. Decision trees
4. Random forest

Each of the method is described in detail in the next section.

Evaluating the classification methods

Confusion Matrix:



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Fig. 5 Confusion matrix

We can compute different metrics from the confusion matrix.

1. Accuracy: Ratio of correct predictions and total predictions

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP}$$

2. Precision: Positive Predictive Value

$$Precision = \frac{TP}{TP + FP}$$

3. Recall: True Positive Rate or Sensitivity

$$Recall = \frac{TP}{TP + FN}$$

Model

All the results in the upcoming section are computed on the test data after training the model using the training data. We have split the data into train-test data in the ratio of 90%-10%.

Training samples = 512 (90% of the total data)
Testing samples = 57 (10% of the total data)

Also, we have made sure that both training and test have almost equal samples of benign and malignant tumor.

Classification methods

A. Logistic regression:

Logistic regression is used when the output is categorical (in our case - benign or malignant). It forms a statistical model which uses the sigmoid function to map the input data to the two output classes. The logistic curve is the common 'S' shape which takes any real value (-∞ to ∞) and yields output between 0 and 1 depicting the probabilities of belonging to a particular class. While training a data using logistic regression model, it tries to find the optimal decision boundary that best separates the two classes i.e. logistic regression splits the feature space linearly. This is binary logistic regression. If the output or the target variable has two or more classes, then the model is called multinomial logistic regression.

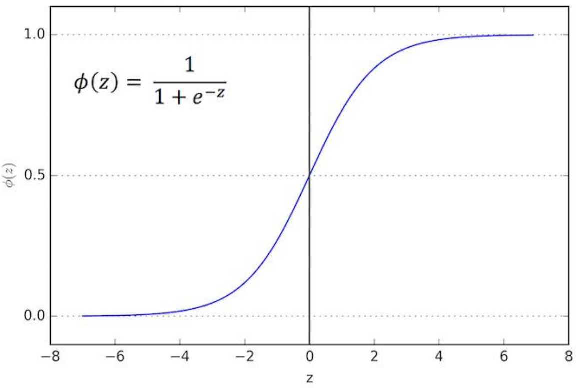
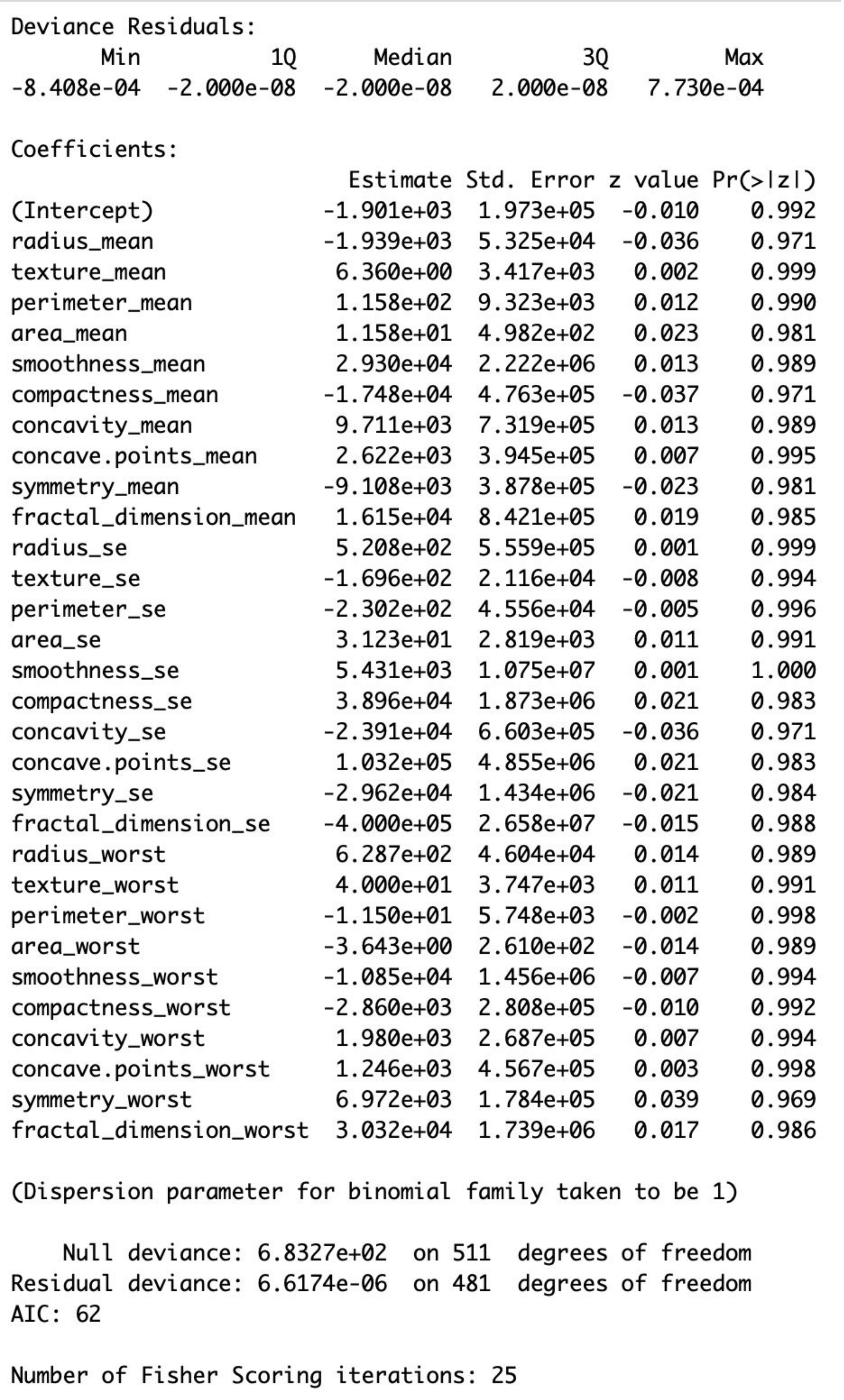


Fig.6 : Sigmoid curve/ Logistic function

Results:

a. *Coefficients for logistic regression:*

The values of the coefficients determine whether the change in an input feature affects the output feature in a positive or negative way and by how much. Interpretation of the parameters in the coefficient table of logistic regression is as follows:



```
Deviance Residuals:
       Min          1Q      Median          3Q         Max
-8.408e-04  -2.000e-08  -2.000e-08   2.000e-08   7.730e-04

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.901e+03  1.973e+05  -0.010    0.992
radius_mean             -1.939e+03  5.325e+04  -0.036    0.971
texture_mean             6.360e+00  3.417e+03   0.002    0.999
perimeter_mean           1.158e+02  9.323e+03   0.012    0.990
area_mean                1.158e+01  4.982e+02   0.023    0.981
smoothness_mean          2.930e+04  2.222e+06   0.013    0.989
compactness_mean        -1.748e+04  4.763e+05  -0.037    0.971
concavity_mean           9.711e+03  7.319e+05   0.013    0.989
concave.points_mean      2.622e+03  3.945e+05   0.007    0.995
symmetry_mean           -9.108e+03  3.878e+05  -0.023    0.981
fractal_dimension_mean   1.615e+04  8.421e+05   0.019    0.985
radius_se                5.208e+02  5.559e+05   0.001    0.999
texture_se              -1.696e+02  2.116e+04  -0.008    0.994
perimeter_se            -2.302e+02  4.556e+04  -0.005    0.996
area_se                  3.123e+01  2.819e+03   0.011    0.991
smoothness_se            5.431e+03  1.075e+07   0.001    1.000
compactness_se           3.896e+04  1.873e+06   0.021    0.983
concavity_se            -2.391e+04  6.603e+05  -0.036    0.971
concave.points_se        1.032e+05  4.855e+06   0.021    0.983
symmetry_se             -2.962e+04  1.434e+06  -0.021    0.984
fractal_dimension_se    -4.000e+05  2.658e+07  -0.015    0.988
radius_worst             6.287e+02  4.604e+04   0.014    0.989
texture_worst            4.000e+01  3.747e+03   0.011    0.991
perimeter_worst         -1.150e+01  5.748e+03  -0.002    0.998
area_worst              -3.643e+00  2.610e+02  -0.014    0.989
smoothness_worst        -1.085e+04  1.456e+06  -0.007    0.994
compactness_worst       -2.860e+03  2.808e+05  -0.010    0.992
concavity_worst          1.980e+03  2.687e+05   0.007    0.994
concave.points_worst     1.246e+03  4.567e+05   0.003    0.998
symmetry_worst           6.972e+03  1.784e+05   0.039    0.969
fractal_dimension_worst  3.032e+04  1.739e+06   0.017    0.986

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.8327e+02  on 511  degrees of freedom
Residual deviance: 6.6174e-06  on 481  degrees of freedom
AIC: 62

Number of Fisher Scoring iterations: 25
```

Fig. 7 Coefficient table of logistic regression for all the 30 features

1. *Standard error:*

The standard error of the coefficients estimates the variability of the coefficients estimation by taking samples from the same population repetitively. This parameter is used to measure the precision of the estimate of the coefficients. Smaller the standard error, less variable is the estimation and thus more precise is the coefficient estimate.

2. *Confidence interval for the coefficients:*

Confidence intervals (considering 95% CI) are the range of values that are likely to contain the true value of the coefficient for each term in the model. The confidence interval is used to assess the practical significance of the results. If the interval is too wide, increasing the sample size of the model for training and testing produces better results.

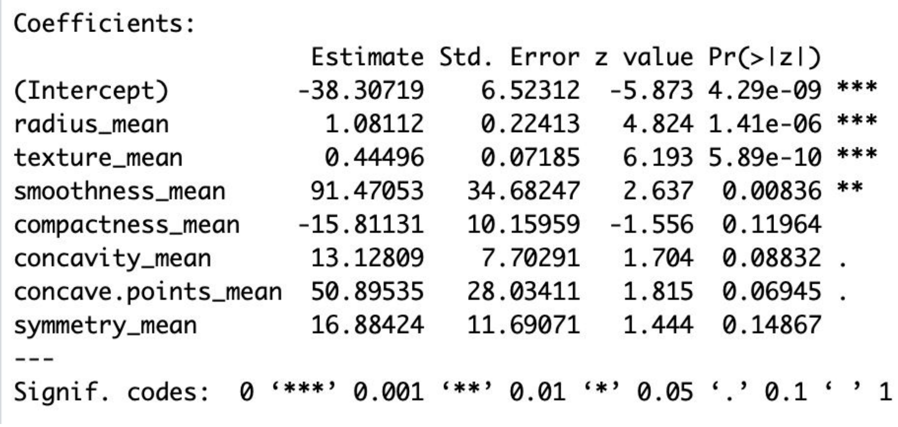
3. *z-value:*

Z-value is a numerical measurement that measures the ratio between the coefficient and its standard error. It is usually used to estimate the coefficient's relationship to the mean of the coefficient. Z-values that are much greater than 0 (either negative or positive) indicate that the coefficient estimate is large and precise enough to be statistically different from 0. Z-values which are closer to 0 depict that the coefficient estimate is too small and imprecise to be sure of the effect of the coefficient on the output.

4. *p-value:*

The p-value is a probability that measures the evidence against the null hypothesis. Lower probabilities provide stronger evidence against the null hypothesis. The null hypothesis is that the term's coefficient is equal to zero, which indicates that there is no association between the term and the response. If the p-value is less than or equal to the significance level (usually 0.05), then it can be inferred that there is a statistically significant association between the input features and the output.

These parameters are used to determine the importance of the features and determine fair estimates of the coefficients. As we can see from the figure-7, the 30 feature predictors show little statistical association significance. To comprehend the feature associations with the output better, we reduce the feature map and then estimate the coefficient parameters. The features for which the estimates satisfy the acceptable criteria of the above parameters are selected in the final model. The following image shows the coefficient table after feature reduction:



```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -38.30719    6.52312  -5.873 4.29e-09 ***
radius_mean           1.08112    0.22413   4.824 1.41e-06 ***
texture_mean          0.44496    0.07185   6.193 5.89e-10 ***
smoothness_mean      91.47053   34.68247   2.637  0.00836 ** 
compactness_mean    -15.81131   10.15959  -1.556  0.11964    
concavity_mean       13.12809    7.70291   1.704  0.08832 .  
concave.points_mean  50.89535   28.03411   1.815  0.06945 .  
symmetry_mean        16.88424   11.69071   1.444  0.14867    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 8 Coefficient table for logistic regression for the features after model reduction by reducing the feature space

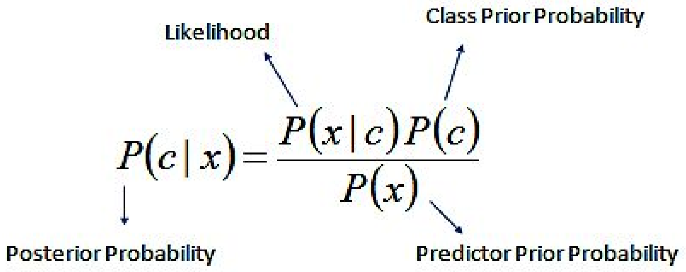
b. Confusion matrix:

	Predicted Benign	**Predicted Malignant**
Actual Benign	37	1
Actual Malignant	6	13

- Accuracy: 0.8772
- Precision: 0.9287
- Recall: 0.6842

B. Naive Bayes classifier:

Naive Bayes classifier is a supervised learning model which is based on the Bayes theorem. This model assumes the conditional independence between every pair of feature. The underlying math behind the model can be formulated in terms of the class *c*, feature space *x* and the bayes theorem as follows:


Likelihood
Class Prior Probability
$P(c\mid x)=\frac{P(x\mid c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability

$$P(c\mid X) = P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

Fig. 9: Bayes Theorem

This independent feature model forms the Naive bayes probability model. To use this model as a classifier, we combine this with a decision rule. One example of the decision rule is to evaluate the above probability estimation for each class *c*, and select the class which has the maximum posterior probability. The denominator can be taken as a constant as it does not play a role in determining the maximum class probability. One of the drawbacks of Naive Bayes model is that it assumes all the features to be conditionally independent. So, if some of the features are in fact dependent on each other (in case of a large feature space), the prediction might be poor.

Results:

a. *Confusion matrix:*

	Predicted Benign	**Predicted Malignant**
Actual Benign	41	1
Actual Malignant	2	13

- Accuracy = 0.947
- Precision = 0.928
- Recall = 0.866

C. Decision Trees

Decision trees form a supervised learning model where the data is continuously split based according to a certain parameter. Here, we consider classification trees for categorical data instead of regression trees. The splitting is done repetitively until the classes are pure i.e. the elements in the class belong to a single output category. This convergence condition and the selection of the splitting variable (feature) is based on parameters defined such as gini impurity, ID3 index, information gain which uses entropy estimation.

Results

1. *Decision tree:*

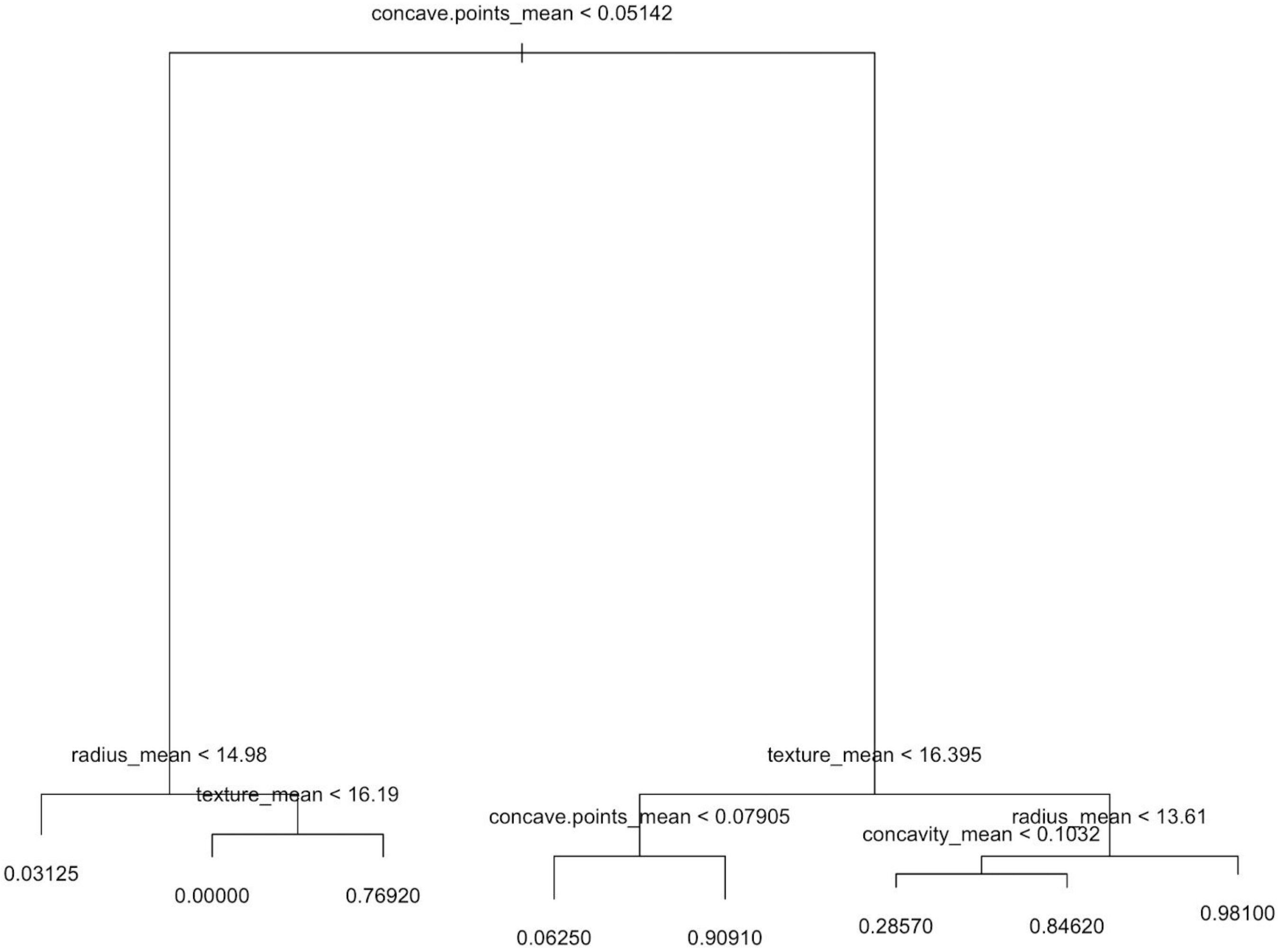


Fig 10: Decision tree determined after training on the data

The feature on the top level of the decision tree is the most heterogeneous feature. If we first split using that, we can separate the classes clearly. Figure 11 shows the distribution of the features, categorized by class. We can see that the distribution of concave.points is very different for Benign and Malignant class, hence its value can be used to classify the data. In the decision tree obtained, concave.points does give the best split. On the other hand, the distribution of fractal dimension dimension tells nothing about the cancer class, and it has been altogether omitted in the decision tree.

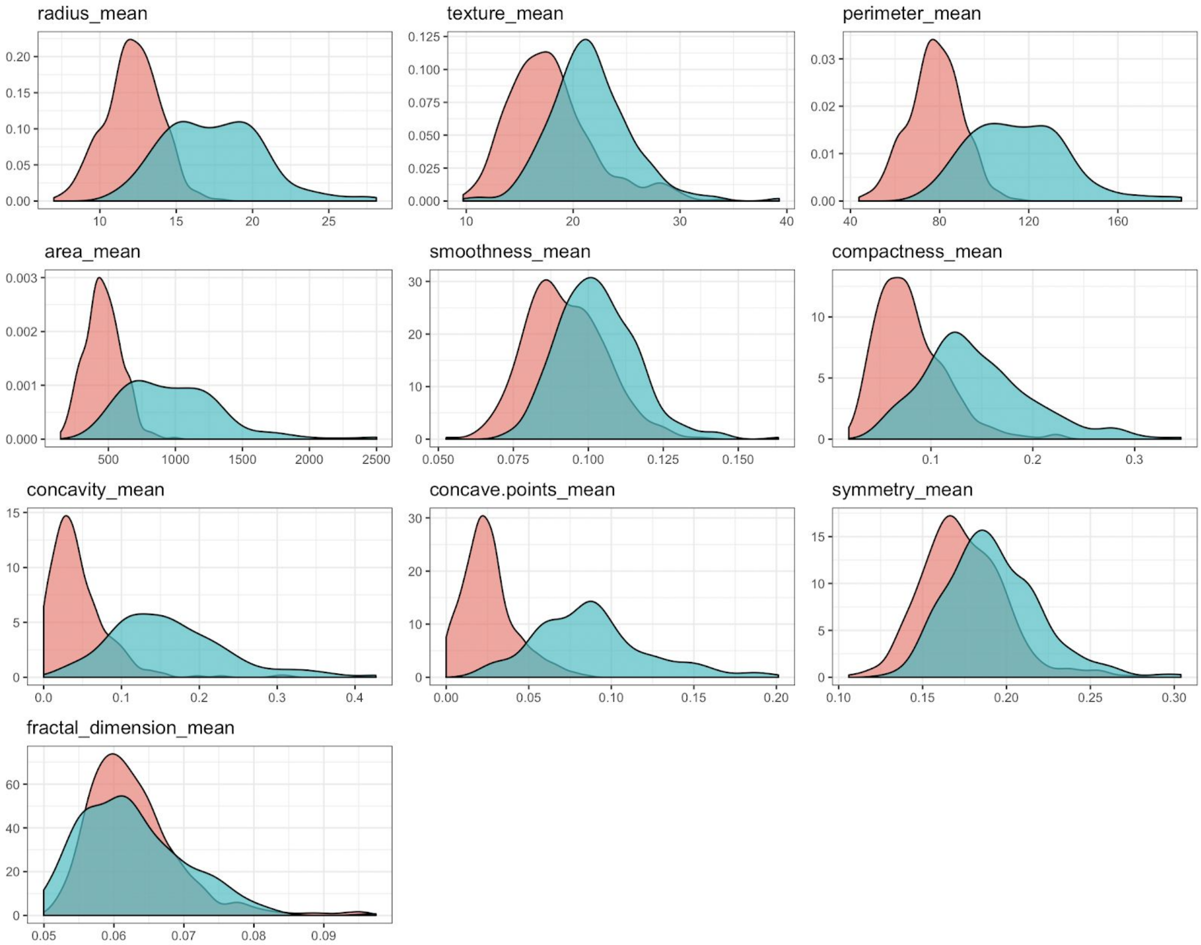


Fig 11: Feature wise and class wise (benign and malignant) histogram plots

2. *Confusion Matrix:*

	Predicted Benign	**Predicted Malignant**
Actual Benign	43	0
Actual Malignant	0	14

- Accuracy = 1
- Precision = 1
- Recall = 1

D. Random Forest:

Random Forest consists of a large number of individual decision trees that operate as an ensemble which is basically using multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Each individual tree in the random forest produces a prediction of a class and the class with the occurs the most from these predictions is our final model's prediction. The key feature behind this model is that the trees have low correlation between them. Uncorrelated models often produce ensemble predictions that are more accurate than any of the individual predictions. This is because the individual trees prevent each other from their individual errors.

Results:

Result on training data

	Predicted Benign	**Predicted Malignant**	**Class Error**
Actual Benign	298	16	0.05095541
Actual Malignant	19	179	0.09595960

Result on test data

	Predicted Benign	**Predicted Malignant**
Actual Benign	43	1
Actual Malignant	0	13

- Accuracy = 0.9825
- Precision = 0.9285
- Recall = 1

Using all the features
Accuracy = 1

Summary:

1. In our case study of detecting cancer, the main aim is to minimize False-Negative rates. Thus, recall is an important parameter to evaluate our models.
2. Logistic regression has low recall values as compared to others as it faces the potential drawback of inability to solve nonlinear problems.
3. Naive Bayes assumes the features to be independent which practically is not the case and hence the probability estimations might not be accurate.
4. Decision trees and Random forests give comparable results. Theoretically, Random forest should give better results than decision trees. However, in our case, a single decision tree is able to capture the importance of features accurately.

Learning Outcome:

- Applied concepts of correlation and PCA for feature space reduction.
- Classifiers like Naive Bayes, which is the direct implication of Bayes rule.
- Other supervised learning based classifiers such as Logistic regression, Decision trees and Random forests.
- Statistical significance of coefficient estimation in logistic regression using the coefficient table and reducing the feature space based on the table.

References:

1. https://towardsdatascience.com/understanding-random-forest-58381e0602d2
2. https://towardsdatascience.com/decision-tree-classification-de64fc4d5aac
3. http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/
4. https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/binary-logistic-regression/interpret-the-results/all-statistics-and-graphs/coefficients/